ČESKÁ ZEMĚDĚLSKÁ UNIVERZITA V PRAZE

PROVOZNĚ EKONOMICKÁ FAKULTA

Datové sady pro trénování umělé inteligence v oblasti fenotypizace rostlin

disertační práce

Autor: RNDr. Alexander Galba

Školitel: doc. Ing. Edita Šilerová, Ph.D.

Katedra informačních technologií

Poděkování

Rád bych touto cestou poděkoval doc. Ing. Editě Šilerové, Ph.D. a doc. Ing. Janu Masnerovi, Ph.D. za jejich odborné vedení. Dále děkuji Mgr. Janě Kholové, Ph.D. za spolupráci na vědeckých projektech a svojí manželce Lucii za podporu.

Datové sady pro trénování umělé inteligence v oblasti fenotypizace rostlin

Abstrakt

Disertační práce se zabývá návrhem, implementací a ověřením metod pro přípravu datových sad určených k trénování nástrojů umělé inteligence v oblasti fenotypizace rostlin, v prostředí vysoce výkonných 3D fenotypizačních platforem. Práce představuje efektivní metodiku pro předzpracování, anotace a distribuci 3D dat rostlin rostoucích v zápoji, u nichž tradiční metody nedosahují požadovaných výsledků z důvodu vysoké scénické komplexity a heterogenity. Práce navrhuje postup pro automatizované odstranění pozadí z multispektrálních 3D skenů, semi-automatizovaný postup anotace dat, včetně systému řízení kvality a nový index komplexity umožňující objektivní rozdělení datových sad i augmentaci 3D dat. Součástí práce je vytvoření a veřejné zpřístupnění anotované 3D datové sady rostlin získané platformou LeasyScan, zahrnující 223 multispektrálních skenů anotovaných na úrovni orgánů. Navržené postupy byly dále využity v aplikaci dynamických bayesovských sítí pro predikci průběhu virové choroby Sterility Mosaic Disease u plodiny Pigeonpea. Výsledky potvrzují, že navržené metody významně zvyšují kvalitu dat, snižují nároky na manuální práci a umožňují spolehlivější trénování modelů umělé inteligence ve fenotypizačních úlohách. Práce přináší ucelenou metodiku zpracování 3D dat a přispívá k rozvoji digitální fenomiky i automatizovaných postupů využitelných v agronomickém i biologickém výzkumu.

Klíčová slova: segmentace pozadí, 3D zobrazování, zpracování pointcloud, strojové učení, fenotypizace rostlin, precizní zemědělství, umělá inteligence, anotace dat, datové sady, 3D skenování

Datasets for training artificial intelligence in plant phenotyping

Abstract

The dissertation deals with the design, implementation and verification of methods for preparing datasets intended for training artificial intelligence tools in the field of plant phenotyping, in the environment of high-performance 3D phenotyping platforms. The main goal is to create an effective methodology for preprocessing, annotation and distribution of 3D data of plants growing in environments where traditional methods do not achieve the desired results due to high scenic complexity and heterogeneity. The work proposes a procedure for automated background removal from multispectral 3D scans, a semi-automated data annotation procedure including a quality management system and a new complexity index enabling objective data set partitioning and 3D data augmentation. The work includes the creation and public disclosure of an annotated 3D plant dataset obtained by the LeasyScan platform, including 223 multispectral scans annotated at the organ level. The proposed procedures were further used in the application of dynamic Bayesian networks for the prediction of the course of the viral disease Sterility Mosaic Disease in Pigeonpea crops. The results confirm that proposed methods show data quality, reduce manual labor requirements and more reliable training of artificial intelligence models in phenotyping tasks. The work brings a comprehensive methodology for processing 3D phenotypic data and lease to the development of digital phenomics and automated procedures usable in agronomic and biological research.

Keywords: background segmentation, 3D imaging, pointcloud processing, machine learning, plant phenotyping, precision agriculture, artificial intelligence, data annotation, datasets, 3D scanning

1 Obsah

2 Úvod

Computer Vision (dále jen CV) je jednou z disciplín umělé inteligence. Jejím cílem je umožnit počítačovým systémům získávat, interpretovat a porozumět informacím obsaženým v digitálních formách obrazu nebo videozáznamu podobně, jako to dělá člověk. Oblast CV se transformovala od jednoduchých geometrických a detekčních úloh ke komplexním modelům. Dnes nástroje CV dokážou segmentovat digitální scény, rozpoznávat objekty, rekonstruovat 3D prostředí či generovat realistická vizuální data. Význam CV je zřetelný v řadě oborů, jako je robotika, automobilový průmysl, medicína, bezpečnost, zemědělství a další.

Zejména v poslední době dochází k zásadnímu posunu využití CV v 3D díky rozvoji metod hlubokého učení. Tyto přístupy umožňují zvýšit výkon v klasifikaci, detekci nebo segmentaci obrazových a prostorových dat. Používání nástrojů CV umožňuje zvyšovat efektivitu výzkumu v různých oblastech.

Jednou z možností multidisciplinárního využití CV představuje oblast, která se zaměřuje na systematické měření, analýzu a interpretaci fenotypických charakteristik organismů. **F**enomika se snaží kvantifikovat výsledný fenotyp, komplexní soubor morfologických, fyziologických a biochemických vlastností organismů v různých časových a prostorových měřítkách. Fenotyp je výsledkem dynamické interakce genotypu s prostředím, a jeho pochopení je přínosem pro predikci biologického chování, optimalizaci růstu a zvýšení odolnosti rostlin vůči různým faktorům.

Fenotypizace představuje metodický rámec fenomiky. Je to soubor postupů a technik, jejichž cílem je měření, popis a kvantifikace fenotypových znaků organismů. Tyto znaky mohou zahrnovat morfologické, fyziologické, biochemické i funkční charakteristiky v různých prostorových a časových měřítkách. Fenotypizace zahrnuje oblast získávání primárních dat, jež fenomika dále analyzuje, integruje a interpretuje.

Z hlediska hierarchického vztahu lze fenotypizaci považovat za podmnožinu fenomiky. Fenomika zahrnuje nejen samotné měření fenotypů, ale také návrh experimentů, správu a integraci rozsáhlých datových souborů, vývoj analytických metod a modelování vztahů mezi fenotypem, genotypem a prostředím. Fenotypizace naproti tomu představuje konkrétní realizaci experimentů a postupů s cílem sběru dat**.**

Vývoj fenomiky je úzce spojen s pokrokem v měřicích, zobrazovacích a analytických technologiích. Tradiční fenotypování rostlin bylo často časově náročné, subjektivní s limitovanou velikostí sledovaných populací. Používání moderních automatizovaných senzorových systémů, 3D skenování, multispektrálních zobrazovacích technik a pokročilých metod pro analýzu obrazu, zásadně transformovalo tuto oblast.

CV přináší do moderní fenomiky přesné a prostorově rozlišené měření rostlinných vlastností, což otevírá nové možnosti pro kvantitativní hodnocení růstu, morfologie a fenologických fází.

CV napomáhá integraci heterogenních datových sad, zvládnutí velkých objemů dat a rozvoji algoritmů, které dokážou robustně extrahovat biologicky relevantní informace. Schopnost analyzovat velké objemy heterogenních dat umožňuje propojit fenotypická data s genetickými a environmentálními informacemi, aby bylo možné odhalit determinanty znaků a předpovědět chování rostlin v reálných podmínkách.

Cílem této disertační práce je přispět k metodickému rozvoji oblasti CV se zaměřením na kvantitativní analýzu komplexních fenotypických znaků a na jejich praktické využití. Práce se soustředí na inovativní metody použití nástrojů CV pro měření fenotypů, využití prostorových dat a aplikaci moderních výpočetních přístupů pro zpracování fenotypických souborů dat. Výsledky disertace poskytují nástroje a metody informačních technologií, které umožní rychlejší a přesnější fenotypování, optimalizaci selekčních procesů a návrh udržitelných zemědělských strategií.

V širším kontextu aplikace nástrojů CV ve fenomice přispívají k řešení globálních výzev, jako jsou zajištění bezpečnosti potravin, adaptace zemědělství na změny klimatu a udržitelné využívání přírodních zdrojů. Tento přístup zdůrazňuje propojení vědeckého výzkumu s jeho praktickými aplikacemi.

3 Výzkumná mezera – Cíle práce

Disertační práce představuje výsledky výzkumu v oblasti metod pro zpracování 3D prostorových dat pro kvantitativní analýzu komplexních fenotypických znaků. Tento výzkum probíhá pod patronací Katedry informačních technologií a ve spolupráci se zahraničními partnery. Práce se zabývá pozorováním rostlin ve venkovním prostředí v zápoji. Tento přístup je odlišný od většiny tradičních metod zkoumajících rostliny jednotlivě. Zvolený přístup umožňuje minimalizovat manipulaci se zkoumanými rostlinami. Cílem je propojení různých vědných oborů pro vytvoření vysoce výkonných metod, které umožňují zvýšit efektivitu výzkumných záměrů, jako šlechtění rostlin nebo adaptace na změny klimatu. Tento multidisciplinární přístup umožňuje zapojení nástrojů umělé inteligence a algoritmů do oblastí výzkumu s vyšším podílem manuálních činností. Při použití vysoce výkonných metod dochází ke kompromisu mezi množstvím dat a jejich kvalitou a tím k dalším vědeckým výzvám.

První část disertační práce se zaměřuje na oblast oddělení pozadí (půdy, nádob, zavlažovacích trubic a dalších) od rostlin ve 3D bodových mračnech. Na základě analýzy dostupných vědeckých článků a publikací, lze tuto problematiku řešit různými metodami. Tyto metody jsou v mnoha případech určené pro jiné domény, například pro automobilový průmysl nebo v rámci odstranění pozadí požadují nižší úroveň přesnosti, přičemž mohou odstraňovat i body náležející pozorovaným objektům, což vede ke ztrátě dat. Použití nástrojů umělé inteligence do metody odstranění pozadí, přineslo požadované výsledky.

Druhá část disertační práce se zaměřila na tvorbu anotované datové sady 3D prostorových dat získaných pomocí vysoce výkonné fenotypizační platformy LeasyScan. Datová sada obsahuje 223 multispektrálních 3D skenů rostlin, anotovaných na úrovni orgánů (děložní listy, listy, řapíky, stonky, celé rostliny), pomocí technologie PlantEye F600. Pro tvorbu těchto datových sad byly použity metody uvedené v disertační práci. Datová sada je zveřejněna ve formě Open Access a umožňuje vývoj AI modelů pro 3D CV rostlin širší komunitě vědců.

Třetí část disertační práce je zaměřena na další kroky v procesu zpracování dat pro použití nástrojů umělé inteligence. To zahrnuje oblast anotace objektů pro potřeby tréninku modelů neuronových sítí. Tento proces je náročný na lidskou práci a čas. To způsobuje relativně menší produkci kvalitně anotovaných dat. Návrh nové metody, který zahrnuje

zapojení vytvořených algoritmů, umožnil urychlení celého procesu. Součástí této části byla i analýza procesu kontroly dat a návrh postupu kontroly.

Čtvrtá část disertační práce se zabývá možností využití 3D zpracovaných dat i v dalších oborech souvisejících s fenomikou, konkrétně se zaměřuje na detekci choroby rostlin. Výzkum zkoumal metody predikce vývoje choroby rostlin formou modelování jejich průběhu i s malým množstvím trénovacích dat. Tyto metody významně snižují nároky na ruční měření a urychlují proces šlechtění.

Téma disertační práce je v souladu s cíli výzkumu probíhajícím na Katedře informačních technologií a zapadá do rámce mezinárodní vědecké komunity zabývající se digitálními fenotypizačními metodami. Výsledky mají potenciál významně rozvinout současné metodické postupy směrem k řešením s vysokou úrovní automatizace, čímž přispějí k dalšímu rozvoji tohoto výzkumného směru.

4 Metodická poznámka

Práce využívá celé spektrum výzkumných metod. Teoretické metody byly využity hlavně v procesu poznávání zvolené problematiky. Analýza byla využita jako hlavní metoda při studiu vědecké literatury a vymezení dílčích oblastí výzkumu a umožňuje rozklad problémů na jednotlivé části a jejich zkoumání. Syntézou získaných poznatků byl pak vytvořen ucelený přehled zkoumané problematiky a vymezeny její zákonitosti. V dílčích oblastech byly využity i další vědecké metody jako porovnání (komparace), analogie a generalizace (Ochrana, 2019). Teoretické metody byly využity především pro specifikaci a vymezení oblasti výzkumu, zjištění stavu vědeckého poznání ve zvolené oblasti a identifikaci výzkumných příležitostí.

Empirické výzkumné metody zahrnovaly především experiment a měření. Jedná se kvantitativní metody, které umožňují srovnávat vlastnosti zkoumaných jevů či objektů a vyjádřit získané poznatky formou dat. Je nutné, aby zkoumané vlastnosti byly konstantní za stejných podmínek a aby získaná data vlastností byla kvantitativně vyjádřitelná pomocí porovnávacích vztahů (Široký, 2011). Tyto metody byly využity především při získávání dat, která tvořila digitální geometrický 3D model pozorovaných rostlin. Tyto experimenty probíhali v kooperaci s pracovištěm The International Crops Research Institute for the Semi-Arid Tropics v Indii (dále jen ICRISAT). Pro dosažení důvěryhodného průběhu experimentů, byla kontrola a řízení podmínek experimentů prováděna vzdáleně z pracoviště Katedry informačních technologií, včetně následných procesů zpracování dat.

Vzhledem k tomu, že součástí výzkumu je také návrh, implementace a ověřování algoritmů, práce používá specifické varianty vědeckých metod vyvinutých v informatice a výpočetních vědách. Tyto přístupy doplňují tradiční hypoteticko-deduktivní rámec o prvky designu artefaktů, experimentální evaluace a iterativní optimalizace. Konkrétněji se jedná o metodologické rámce Design Science Research, Computational Science a informatickou (algoritmickou) vědeckou metodu.

Design Science Research (Wieringa, 2014) je vědecká metodologie zaměřená na tvorbu nových artefaktů, které řeší existující problém. Artefaktem se v kontextu informatiky rozumí algoritmus, metoda, softwarová komponenta, framework, datová struktura nebo systém. Design Science Research vychází z premisy, že vytváření inovativních technických řešení představuje plnohodnotnou formu vědeckého poznání.

Computational Science je vědecký přístup, který využívá výpočetní modely a algoritmy k pochopení, simulaci nebo analýze složitých jevů. V přírodních vědách je algoritmus prostředkem k získání nových poznatků, nikoli cílem samotným (Shiflet and Shiflet, 2014).

Třetí metodologickou součástí je informatická vědecká metoda, která přímo staví algoritmus do centra vědeckého zkoumání. Tento přístup je založen na předpokladu, že nové algoritmy a jejich vlastnosti tvoří samostatný vědecký přínos, bez ohledu na konkrétní doménu aplikace.

4.1 Rámcový metodický postup

Metodika disertační práce byla transformována do praktického postupu disertační práce, který lze shrnout do následujících bodů:

- Přehled literatury, analýza dostupných vědeckých informačních zdrojů s cílem zpracování přehledu současného stavu řešené problematiky, identifikace a vymezení výzkumné mezery.
- Studium informací výrobce 3D skeneru společnosti Phenospex, včetně návštěvy společnosti v jejím sídle v Holandsku a konzultace s vývojovým týmem.
- Sestavení mezinárodního týmu, stanovení komunikačních kanálů, pravidelných schůzek a projektových parametrů výzkumu.
- Formulace pracovních hypotéz výzkumu, zúžení problematiky za účelem vymezení vhodného rozsahu výzkumu.
- Nastavení a kontrola procesu pořizování dat z pracoviště ICRISAT v Indii.
- Tvorba algoritmů a jejich ověřování pro zpracování dat.
- Nastavení experimentů a ověřování navrhovaných postupů.
- Experimentální ověření pracovních hypotéz formou praktické implementace navržených metodických postupů.
- Prezentace dílčích částí výzkumu formou publikací ve vědeckých periodikách a na odborných konferencích.
- Celkové vyhodnocení zvolených postupů a výsledků výzkumu.
- Generalizace získaných poznatků a využití metodiky v rámci navazujícího výzkumu.

5 Literární přehled – současný stav výzkumu

Analýza současného stavu výzkumu se zaměřila na oblasti CV, zpracování dat, fenomiky a fenotypizace. Dále byly zkoumány neinvazivní metody pozorování rostlin, vysoce výkonné platformy pro neinvazivní pozorování rostlin, přípravu dat pro použití nástrojů umělé inteligence, metody augmentace a další.

5.1 Zpracování dat pro použití nástrojů umělé inteligence

5.1.1 Předzpracování dat

Pojem předzpracování dat zahrnuje různé, velice často proprietární metody, které extrahují a převádějí data získaná z různých zařízení do požadované formalizované podoby. Výrobci různých zařízení generují data často ve svých formátech, včetně metadat a dat nesouvisejících přímo s pozorovanou realitou, jako například data o průběhu. Pro takové zpracování dat se používají metody přizpůsobené konkrétnímu zařízení. Důležitá je volba výsledného formátu a obsahu dat, aby nedošlo k překážkám při dalším zpracování dat pokročilými metodami. Pro 3D data jsou používané formáty dat jako Polygon File Format ("PLY (file format)," 2025) nebo Point Cloud data format (Wang et al., 2020). Dále je nutné, v co nejvyšší míře zachovat původní hodnotu dat, která může být ovlivněna různými transformacemi. Při takovém postupu je nutné komunikovat s výrobcem zařízení nebo mít k dispozici kompletní technickou dokumentaci.

5.1.2 Anotace dat

Proces anotace dat je časově nejnáročnější fází zpracování dat. Tento proces zahrnuje lidskou činnost, kdy dochází k anotaci dat ve smyslu významu objektů a jejich částí, které data představují. Tento proces zahrnuje také definici kolekce objektů, které budou anotovány. Pro tento proces existují již platformy, které celý postup usnadňují ("Amazon Mechanical Turk," ; "Multi-sensor data labeling platform for robotics & AV | Segments.ai,"; "Supervisely: Curate, Label and Build Production Models in One Platform,"). Tyto platformy nabízejí i sdílení již anotovaných dat. Tímto způsobem lze znásobit množství dat, pokud se shoduje typ a účel použití dat.

Kompromisem pro data pořízená vysoce výkonnými metodami je často nižší kvalita dat. To způsobuje náročnější proces anotace a generuje se tím menší množství kvalitně

anotovaných dat. Při anotaci dat z oblasti fenotypizačních úloh, je pro proces anotace nutné vybrat lidi, kteří rozeznávají rostliny a jejich části pro různé odrůdy. Tím se tento proces liší od procesu anotace dat, například z prostředí automobilového průmyslu, kde anotované objekty reprezentují auta, lidi a další všeobecně známé objekty.

5.1.3 Augmentace

Augmentace je postup, při kterém se z již existujících anotovaných dat generují další označená data. Metody augmentace lze rozdělit do různých kategorií na základě zvolené perspektivy. V závislosti na tom, zda se augmentace týká celého objektu nebo jeho části, mohou být metody globální a lokální (Hahner et al., 2022). Pokud se zaměříme na kontext metod augmentace, můžeme je rozdělit na bodově orientované a objektově orientované. Pokud vezmeme v úvahu složitost, můžeme augmentace rozdělit na základní a pokročilé. Mezi standardní geometrické augmentace patří rotace, translace a škálování (Qiu et al., 2021). Pomocí analytické geometrie můžeme aplikovat další augmentace, jako je převrácení, zrcadlení, elastické zkreslení a vážená lokální transformace (Kim et al., 2021; Wu et al., 2022). Mezi metody augmentace patří také uměle generované datové sady pomocí generativního softwaru, umělé inteligence nebo rekurzivních gramatických metod (Ma et al., 2020). Výzkum zaměřený na rozpoznávání objektů v robotice nebo automobilovém průmyslu pracuje s 3D mračny bodů, kde hloubka a celková struktura detekovaných objektů nejsou často podstatné, např. detekce povrchu nebo překážky. V těchto případech jsou důležité kontury a vzdálenost. Těmto cílům jsou přizpůsobeny i metody augmentace (Weon et al., 2020). Důležitý je účel, pro který je augmentace provedena. Je nezbytné, aby proces augmentace zachoval anotační informace, jinak by se proces anotace musel opakovat. Konečným cílem celého procesu augmentace je zlepšení výsledků. Tabulka 1 poskytuje přehled metod augmentace. Jsou rozděleny do několika skupin podle typu augmentační transformace. Metody jsou stručně popsány.

Augmentační metoda	Stručný popis	Zdroj
point-stable transformationrotation, translation, scaling, mirroring, flipping, resizing, deformation, elastic distortion, mixup	geometric transformations preserving the number of original points, mostly well and simply algorithmically applicable	(Alomar et al., 2023; Kim et al., 2021; Nekrasov et al., 2021; Wu et al., 2022)
point-nonstable transformation, cropping, dropping, downsampling, jittering, blurring, cutmix, cut-paste,	point-oriented transformations, changes shape parameters, more demanding for application	(Chen et al., 2020; Nekrasov et al., 2021; Zhu et al., 2024)
photometric transformation, color jitter, motion blur, optical noise, color filtering	transformations working with photometric effects, color artifacts	(Cheong et al., 2024; Sheshappanavar et al., 2021; Zini et al., 2023))
object-oriented transformation, GT-sampling	adds additional training objects to scenes; balances classes and increases variety	(Šebek et al., 2022; Shi et al., 2023)
advanced transformations, differentiable automatic data augmentation, deep autoaugment, sample-adaptive,	transformation using advanced data processing methods	(Li et al., 2020; Zheng et al., 2021)
synthetic sample generation, computer-aided design (CAD), recursive grammar, differential neural rendering	methods producing artificially created samples require a separate research process	(Ma et al., 2020; Niemeyer et al., 2020)

Tabulka 1- přehled metod augmentace

5.1.4 Distribuce dat

Při použití nástrojů strojového učení je potřeba modely umělé inteligence natrénovat. Pro tento proces data rozděluji do tří skupin: učící, kontrolní a testovací (train, validation, test). Z důvodu vyhodnocení a použití kontrolních metrik jsou použitá data anotovaná. Pokud je k dispozici dostatečná sada anotovaných dat, je rozdělení prováděno náhodně. Může dojít i k situaci, že se počet dat snižuje nebo zvyšuje, z důvodu rovnoměrného rozdělení. V takovém případě se některá data vynechají nebo duplikují.

Jiná situace nastává, pokud je dat méně. V tomto případě je učení závislé na vhodném rozdělení dat. Náhodné rozdělení nemusí v tomto případě přinášet nejlepší výsledky. Cílem je vytvořit homogenní sady dat. Toto rozdělení je pak závislé na kvalitě, množství a obsahu dat. Jednou z používaných metod je stratifikovaný výběr. Cílem stratifikovaného výběru je vytvořit disjunktní, homogenní skupiny dat na základě hodnoty cílové proměnné. V úlohách učení s více označeními, kde existuje více cílových proměnných, však není jasné, jak by měl být stratifikovaný výběr proveden (Sechidis et al., 2011). V těchto případech se často

vytvářejí kombinace cílových proměnných a jejich hodnot, nebo se konstruují nové proměnné vhodné pouze pro účely rozdělení.

5.2 Fenotypizace a fenomika

Fenomika a fenotypizace tvoří metodologický rámec moderního výzkumu rostlin. Fenotypizace označuje proces měření, kvantifikace a popisu fenotypových znaků organismu, jeho pozorovatelných vlastností, které vznikají interakcí genotypu s prostředím. Fenotypizace je primárně metodickým a technickým postupem, jehož cílem je získat přesné, reprodukovatelné a biologicky relevantní údaje o růstu, morfologii, fyziologii či architektuře rostlin.

Fenomika je širší výzkumná disciplína, jejím cílem je komplexně porozumět dynamice rostlinných systémů, jejich proměnlivosti a reakci na různé biotické a abiotické podněty. Fenomika propojuje fenotypová měření s pokročilými zobrazovacími technologiemi, statistickými modely a výpočetními metodami (např. 3D rekonstrukcí, segmentací nebo modelováním architektury rostlin). Výsledkem je hlubší vhled do toho, jak se fenotyp mění v čase, jak reaguje na prostředí a jak lze tyto změny interpretovat ve vztahu ke genetickým informacím (Kumar et al., 2015).

Fenotypizace také představuje proces sběru dat, fenomika tvoří rámec, který tato data integruje, interpretuje a využívá ke studiu komplexních biologických souvislostí. Fenomiku lze chápat jako disciplínu, jež zahrnuje nejen samotné měření, ale také technologické, analytické a teoretické aspekty popisu fenotypové variability na různých prostorových a časových škálách. Praktická fenomika vyžaduje systematickou a škálovatelnou fenotypizaci, často s využitím automatizovaných a vysokokapacitních metod, které jsou základním předpokladem pro zpracování rozsáhlých datových souborů.

Přesné a správné měření znaků hraje důležitou roli v genetickém vylepšování plodin. Proto v nedávné minulosti došlo v oblasti fenomiky k velkému vývoji. Vyvinuly se jak přímé, tak i zpětné fenomiky (Ndlovu, 2020), které mohou pomoci identifikovat buď nejlepší genotyp s požadovanými znaky, nebo mechanismus a geny, které činí genotyp nejlepším. To zahrnuje vývoj vysoce výkonných neinvazivních zobrazovacích technologií. Tyto technologie umožňují pomocí spekter světla barevně zobrazovat biomasu, strukturu

rostlin, detekovat zdraví listů (chloróza, nekróza), měřit obsah vody v tkáních rostlin a v půdě, měřit teplotu korun/listů a dalších atributů. Tyto fenomické nástroje a techniky umožňují využít potenciálu genomických zdrojů v genetickém vylepšování plodin. Obecně lze ve výzkumu rozlišit dva hlavní směry: simulační a experimentální.

Modely simulace plodin založené na procesech CSM (crop simulation model) jsou dynamické výpočetní nástroje, které simulují vývoj a růst plodiny ve vztahu k podmínkám prostředí jako jsou teplota vzduchu, koncentrace půdní vody, odpařování a koncentrace CO2 v atmosféře, v kombinaci s postupy hospodaření, jako je datum setí, aplikace dusíkatých hnojiv a zavlažování. V současnosti je k dispozici více modelů, včetně volně dostupných, jako například APSIM, DAISY, DSSAT a další (Gavasso-Rita et al., 2024). Dalším směrem simulace je 3D modelování plodin s popisem morfologie a formy rostlin, známé jako funkčně-strukturální modely rostlin FSPM. Na rozdíl od CSM, které obecně modelují procesy na úrovni porostu, FSPM zohledňují procesy na úrovni jednotlivých rostlin, z nichž vycházejí komplexní vlastnosti na úrovni rostlinného společenstva. Různé FSPM se zaměřují buď na nadzemní nebo podzemní části rostliny. Většina z nich nezohledňuje celý životní cyklus rostliny, ale pouze omezený počet procesů v závislosti na řešených výzkumných otázkách. Šíře procesů zohledňovaných v FSPM se však v posledních letech výrazně rozšířila, což jim nyní umožňuje řešit otázky týkající se interakcí plodin s prostředím a managementem, které dříve mohly řešit pouze CSM (Evers et al., 2019). CSM a FSPM se doplňují a mohou se vzájemně informovat o procesech, které probíhají v různých měřítkách. V tomto speciálním případě jsou zvažovány oba typy modelů.

V experimentálním výzkumu se pozoruje vývoj a vlastnosti rostlin v reálném prostředí. Pozorované atributy rostlin jako například biomasa, počet orgánů rostlin a další, se v minulosti zjišťovali manuálně a destruktivním způsobem. Příchodem nových technologií, a hlavně možností automatického rozpoznávaní obrazu a 3D modelů, se těžiště výzkumu přesunulo k nedestruktivním metodám (Muhammad et al., 2025).

5.3 Nedestruktivní metody pozorování rostlin

Nedestruktivní přístupy nabízejí rychlé a spolehlivé vyhodnocení pozorovaných vlastností rostlin a plodin v reálném čase a zároveň zachovávají integritu vzorků.

Nedestruktivní metody se v závislosti na cíle výzkumu zaměřují, buď na konkrétní pozorovanou veličinu, jako například fluorescenci chlorofylu (Dong et al., 2020), měření

výměny plynů (Pflüger et al., 2024), elektrická impedanční tomografie (Basak and Wahid, 2022) a jiné, nebo na pozorování vývoje rostlin v čase prostřednictvím zařízení, které zachycují 2D či 3D, RGB, multispektrální nebo hyper-spektrální odraz rostlin. Zatímco 2D zobrazovací metody jsou zkoumány již po dlouhou dobu, využití 3D zobrazovacích technologií, například LiDARu (Behroozpour et al., 2017) nebo 3D skenování, se dosud rozvíjí v kratším časovém rámci. To platí zejména pro novou generaci senzorů, které dokážou současně zachytit tvar, barvu i další spektrální informace. Pořízená data se následně analyzují a vyhodnocují pomocí nástrojů umělé inteligence. V oblasti rozpoznávání obrazu a 3D modelů jde především o neuronové sítě (Patel et al., 2021; Schindelin et al., 2012; Zhou and Luo, 2009).

Pro oblast 3D skenování rostlin využívá výzkum různé typy zařízení a s nimi spjaté metody. Velká část výzkumných projektů využívá zařízení, která pracují v uzavřených prostorách a skenují rostliny jednotlivě. To sice generuje bohaté 3D modely rostlin, ale zároveň to vyžaduje vyšší míru manipulace s rostlinami a snižuje se tak celková efektivita metody. Řešením jsou systémy minimalizující manipulaci s rostlinami, které se obecně zahrnují pod pojem vysoce propustné/výkonné metody (dále jen HTPM) (Gill et al., 2022). Vysoce výkonná fenotypizace rostlin je pokročilý přístup k rychlému a nedestruktivnímu měření široké škály znaků rostlin ve velkém měřítku s využitím automatizovaných a přesných technologií. Hlavním cílem je urychlit výzkum a šlechtění rostlin generováním velkých datových sad fenotypových znaků, které doplňují genomické informace. Systémy HTPM obvykle integrují různé senzory, zobrazovací technologie a automatizované datové kanály pro pozorování a kvantifikaci růstu, morfologie, fyziologie a reakce rostlin na faktory prostředí. Tyto systémy mohou fungovat v kontrolovaném prostředí (např. skleníky) nebo v terénu a zachycovat znaky, jako je biomasa, architektura porostu, fotosyntetická účinnost, stav vody a stresové reakce v průběhu času (Guo et al., 2023). Tyto metody kromě svých výhod přinášejí také další vědecké výzvy.

Vysoce výkonné fenotypování umožňuje rozsáhlý sběr snímků rostlin a dat ze senzorů ve sklenících a na polích (McCormick et al., 2016; Xiong et al., 2017). Snímky tisíců plodin na polích lze pořizovat současně a nepřetržitě po celou dobu jejich růstu umístěním kamer nebo 3D skenerů. Pro využití těchto dat pro analýzu rostlinných fenotypů je nutné následné zpracování snímků pro extrakci znaků. Segmentace rostlinných objektů je základním krokem v extrakci znaků rostlin (Ge et al., 2016; Ghanem et al., 2015; Guo et al., 2018). Pro

zpracování snímků rostlin, lze použít již existující nástroje jako Leaf-GP (Zhou et al., 2017), Scanalyzer (Paauw et al., 2024) nebo CropSight (Liu et al., 2024) a další. Tyto nástroje obvykle používají pro segmentaci rostlin metody, které fungují dobře pro snímky ze skleníků s homogenním pozadím. Nedokážou však produkovat uspokojivé výsledky segmentace rostlin u snímků se složitým pozadím, zejména u snímků rostlin na poli. Metody neuronových sítí, dokážou přesněji segmentovat snímky rostlin se šumem na pozadí. Tyto metody jsou však založeny na řízeném učení s anotovanými daty. Příprava dostatečně velké trénovací datové sady je časově náročná a pracná.

6 Výsledky

Disertační práce se zaměřila na výzkum v oblasti použití nástrojů umělé inteligence ve fenotypizačních úlohách ve venkovních podmínkách pro rostliny rostoucí v zápoji. Pro výzkum bylo využíváno zařízení umožňující 3D skenování rostlin. Tento způsob umožňuje pořizovat 3D modely rostlin s minimální nutností manipulace a vyšší efektivitou celého procesu. Zároveň generuje velké množství dat a přináší nové vědecké výzvy, například odstranění pozadí, segmentace, detekce objektů a další.

Disertační práce řeší tu část výzkumu, která zpracovává získaná data z experimentů pro jejich další použití nástroji umělé inteligence. To v sobě zahrnuje navržení metod a postupů pro extrakci dat z použitého zařízení (3D skeneru), jejich transformaci do požadovaného formátu a následné zpracování.

Dosažené výsledky zahrnují návrh a ověření metody odstranění pozadí, dat, které nesouvisejí s pozorovanými objekty. Dále implementaci semi-automatického postupu anotace dat, který zahrnuje i systém řízení kvality, generování dat a jejich publikování ve standardizovaném formátu pro možnost dalšího výzkumu a využití dat pro metody modelování průběhu chorob rostlin.

Schéma publikací, které reprezentují výsledky výzkumu je znázorněn na Obrázek 1.

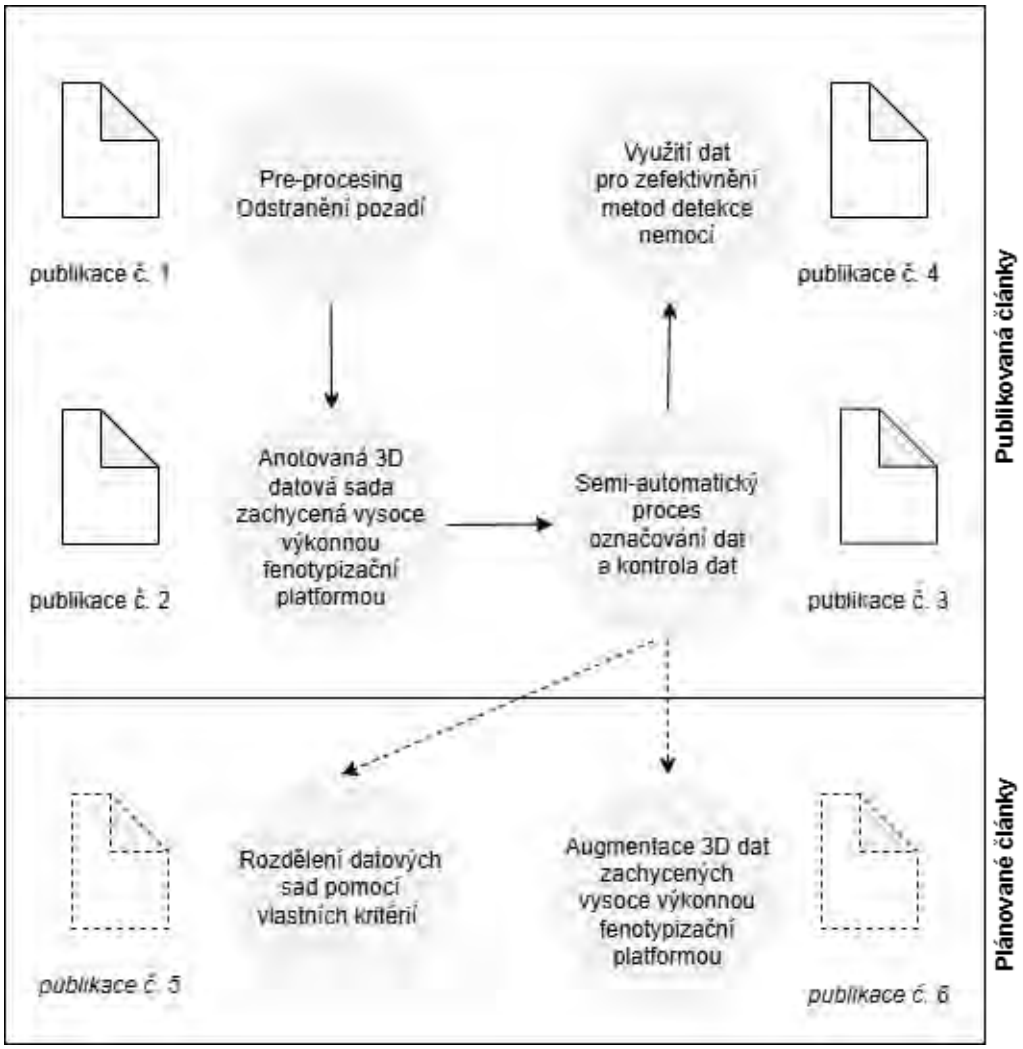
V publikaci č.1 se autor zabývá především oblastí přípravy dat (pre-processing). V publikaci č.2 a č.3 autor řeší hlavní témata publikací, přípravu unikátní anotované datové sady, semi-automatizaci označovacího procesu a kontrolu kvality dat. V publikaci č.4 je řešena příprava dat.

V připravovaných publikacích č. 5 a č.6 autor řeší hlavní ideu vyjádřenou complexity indexem a jeho následné použití v oblasti distribuce dat.

Všechny uvedené publikace využívají poznatky z autorova výzkumu, který je zaměřen na zdrojová data, jejich zpracování a analýzu.

6.1 Výchozí situace a průběh výzkumu

V rámci mezinárodního výzkumného týmu byl zahájen projekt, který propojuje Katedru informačních technologií s výzkumným pracovištěm ICRISAT. Pracoviště ICRISAT disponuje zařízením, které umožňuje 3D skenování rostlin ve venkovním prostředí. Tým měl vzdálenou kontrolu nad zařízením a také kompletní přístup k originálním datům. Na základě těchto výchozích podmínek se realizoval výzkum, který nadále pokračuje a jehož výstupy jsou součástí této disertační práce.



Obrázek 1- schéma postupu a publikací disertační práce

Na schématu Obrázek 1 je znázorněn průběh disertační práce, včetně výstupů, které korelují s výzkumem. Schéma obsahuje jak publikované, tak i aktuálně rozpracované a plánované publikace.

Data použitá ve výzkumu byla generována pomocí platformy LeasyScan (Vadez et al., 2015) s použitím PlantEye F600 a 3D skeneru s multispektrálním zobrazováním. LeasyScan využívá systém duálního skenován, který generuje 3D bodový model ve formě dvou překrývajících se pootočených souborů dat ve formátu .ply ("PLY - Polygon File Format,"). 3D bodový model zachycuje nejen body náležící rostlinám, ale i body patřící okolí, které tvoří body pěstících nádob (květináčů), zeminy a dalšího okolí. Celá oblast je rozdělena do 4*18=72 sektorů a odpovídá jednotlivým pěstícím nádobám, které jsou automatizovaně identifikovatelné prostřednictvím čárového kódu.

Pro další využití dat bylo nejprve nutné je sjednotit. K tomuto účelu byly vytvořeny algoritmy, které tyto operace automatizovaly. Dalším krokem bylo odstranění bodů, jež nepatří pozorovaným rostlinám. Pro tento účel použil výrobce zařízení efektivní a jednoduchý postup, a to odstranění bodů na základě jejich výšky nad zemí. Byla použita známá výška okraje pěstební nádoby. Tato metoda je jednoduše aplikovatelná. Její nevýhodou je, že v praxi při experimentech bývá okraj pěstební nádoby až 10 cm nad povrchem zeminy, což vede ke ztrátě bodů rostlin, zejména v počátcích jejich růstu. Další zkoumanou metodou bylo využití informací o barvě jednotlivých bodů. Tato metoda nebyla nakonec zvolena, vzhledem k neuspokojivým výsledkům. Jako vhodná metoda pro odstranění pozadí byla aplikace neuronové sítě pro segmentaci pozadí. Výstupem této části výzkumu je publikace č. 1.

Publikace č. 1

typ:	vědecký článek
název:	AI-Driven Background Segmentation for HighThroughput 3D Plant Scans
autoři:	Serkan Kartal, Jan Masner, Jana Kholová, **Alexander Galba**, Tharanya Murugesan, Rekha Baddam, Vojtěch Mikeš and Eva Kánská
rok:	2025
vydáno v:	IEEE Access
indexováno:	Web of Science, Impact Factor 3.6, SCOPUS SJR 0.849
kvartil	Q2 JIF, Q1 SJR, Q1 AIS
AIS	0,670
odkaz:	https://doi.org/10.1109/ACCESS.2025.3594406

Přehled literatury ukázal, že není mnoho veřejných datových sad, které poskytují anotovaná data ve formátu 3D bodů. Pokud se ve veřejných repositářích objevují anotovaná data ve formátu 3D, jsou převážně generovaná technologií LiDAR, nebo poskytují data

generovaná vysoce přesnými systémy. Žádná veřejně dostupná datová sada ve formátu 3D bodů generována HTPM nebyla dohledána. V rámci výzkumu byla vygenerována a zveřejněna anotovaná datová sada ve formátu 3D bodů. Tato datová sada představuje anotované 3D mračna bodů rostlin generované HTPM. Zaměřuje se na druhy širokolistých bobovitých rostlin (fazole mungo, fazole obecné, vigna a fazole lima). Datová sada zahrnuje 223 souborů, které poskytují podrobné anotace segmentace na úrovni orgánů pro embryonální listy, listy, řapíky, stonky a celé rostliny. Datová sada nabízí základnu anotovaných dat na podporu vývoje modelů umělé inteligence v 3D počítačovém vidění. Datová sada, kód pro předzpracování, anotace a informační záznam kompatibilní s MIAPPE jsou také součástí a vše je poskytnuto v repositáři GitHub pro další aktualizace a rozšíření. Výstupem této části výzkumu je publikace č.2.

Publikace č. 2

typ:	vědecký článek
název:	Annotated 3D Point Cloud Dataset of Broad-Leaf Legumes Captured by High-Throughput Phenotyping Platform
autoři:	**Alexander Galba**, Jan Masner, Jana Kholová, Serkan Kartal, Michal Stočes, Vojtěch Mikeš, Pavel Šimek, Štěpánka Prokopová, René Fiala, Thorsten Karrer, András Tóth
rok:	2025
vydáno v:	Nature/*Scientific Data*. 2025, vol. 12, no. 1.
indexováno:	Web of Science, Impact Factor 6.9, SCOPUS SJR 1.867
kvartil	Q1 JIF,Q1 SJR, Q1 AIS
AIS	2,661
odkaz:	https://doi.org/10.1038/s41597-025-06049-7

Dalším krokem výzkumu byl proces anotace dat pro potřeby natrénování neuronové sítě s cílem detekce zvolených objektů, rostlin a jejich částí. Tento proces je zdlouhavý a časově náročný. Neuronovou síť při rozpoznávání dat lze použít dvěma způsoby, buď pro detekování bodů podle náležitosti k danému typu objektu – segmentace, nebo pro detekci objektů, včetně vyjádření jejich počtu – detekce objektů. Pro každý přístup se používá jiný způsob anotace dat. Pro segmentaci se vyznačují konkrétní body, pro detekci objektů se množina bodů ohraničí hranolem. Pro urychlení celého procesu byl vyvinut semi-automatizovaný postup, kdy byly označovány pouze body a hranoly byly generovány prostřednictvím algoritmu. Zároveň byl vytvořen nový způsob záznamu dat o anotaci, a to vytvořením metadat ve formátu .csv, který uchovává data o příslušnosti jednotlivých části

rostlin k dané rostlině. Součástí výzkumu byl návrh metody pro kontrolu kvality procesu anotace dat. Výstupem této části výzkumu je publikace č. 3.

Publikace č. 3

typ:	vědecký článek
název:	Application of Quality Management System in the Research Process: A Case Study for Plant Phenotyping Research
autoři:	**Galba, A.**, Kánská, E., Mikeš, V., Vaněk, J. and Jarolímek, J.
rok:	2024
vydáno v:	Agris on-line
indexováno:	Web of Science, Impact Factor 0.7, SCOPUS SJR 0.23
kvartil	Q3 SJR
AIS	N/A
odkaz:	https://doi.org/10.7160/aol.2024.160406

Jednou z významných oblastí fenomiky je i modelování průběhu chorob rostlin. Tato oblast výzkumu obvykle vyžaduje velké množství manuální lidské činnosti, především v oblasti získávání dat. Generování vysoce kvalitních anotovaných dat ve formátu 3D otevřelo možnost zkoumání metod, které využívají matematické nástroje a tím docílilo významného urychlení celého procesu sledování průběhu chorob rostlin. V rámci realizovaných projektů v kooperaci s University of Strathclyde v Glasgow a University of Pisa v Itálii, byl výzkum rozšířen o predikci virového onemocnění Sterility Mosaic Disease u plodiny pigeonpea (Cajanus cajan) s využitím HTPM a dynamických bayesovských sítí. Výstupem této části výzkumu je publikace č.4.

Publikace č. 4

typ:	příspěvek na konferenci
název:	Forecasting Sterility Mosaic Disease in Pigeonpea Using Dynamic Bayesian Networks and 3D Point Cloud High-throughput Scanning Platform
autoři:	Vojtěch Mikeš, Alexander Kocian, Jana Kholová, Jan Masner, Adam Kleczkowski, Mamta Sharma, Stefano Chessa, **Alexander Galba**, Pavel Šimek
rok:	2025
vydáno v:	IEEE Xplore
indexováno:	Web of Science, Impact Factor 3.6, SCOPUS SJR 0.849
kvartil	Q2 JIF, Q1 SJR, Q1 AIS
AIS	0,670
odkaz:	https://doi.org/10.1109/IE64880.2025.11130066

6.2 AI-Driven Background Segmentation for HighThroughput 3D Plant Scans

KARTAL, Serkan; MASNER, Jan; KHOLOVÁ, Jana; GALBA, Alexander; MURUGESAN, Tharanya et al. AI-Driven Background Segmentation for High-Throughput 3D Plant Scans. Online. *IEEE Access*. 2025, roč. 13, s. 136027-136037. ISSN 2169-3536. Dostupné z: https://doi.org/10.1109/access.2025.3594406. [cit. 2025-11-08].

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.Doi Number

AI-Driven Background Segmentation for High-Throughput 3D Plant Scans

Serkan Kartal[1], Jan Masner[2], Jana Kholová[2], Alexander Galba[2], Tharanya Murugesan[3], Rekha Baddam[3], Vojtěch Mikeš[2] and Eva Kánská[2]

[1] Department of Computer Engineering, Çukurova University, 01380, Adana, Türkiye
[2] Department of Information Technologies, Czech University of Life Sciences Prague, 165 00 Prague, Czech Republic
[3] Crop Physiology and Modeling, International Crops Research Institute for the Semi-Arid Tropics (ICRI-SAT), Patancheru, 502 324, Telangana, India

Corresponding author: J. Masner (e-mail: masner@pef.czu.cz).

The results and knowledge included herein have been obtained owing to support from the EC's Horizon Europe funding in the project CODECS, grant no. 101060179, the institutional grant of the Internal grant agency of the Faculty of Economics and Management, Czech University of Life Sciences Prague, grant no. IGA 2023B0005, and the project: "Precizní zemědělství a digitalizace v ČR" (Precision Agriculture and Digitalization in Czech Republic), reg. No. QK23020058.

ABSTRACT Accurate background segmentation in 3D plant phenotyping is crucial for reliable trait assessment but remains challenging. Current methods are either excessively complex, developed for a different domain, or lead to data loss (coordinate-based). This paper addresses these issues by introducing an AI-driven approach using a Multi-Layer Perceptron (MLP) model, leveraging RGB, spatial (XYZ), and near-infrared (NIR) data to enhance precision. The method was evaluated on high-throughput phenotyping data, achieving a classification accuracy of 0.9993, significantly reducing false positives and false negatives compared to coordinate-based segmentation. The proposed approach improved segmentation, particularly in early growth stages and for prostrate species, where traditional methods often fail. The model's impact on leaf area estimation was validated against destructive measurements, demonstrating substantial accuracy improvements, especially for species with small and prostrate canopies. Additionally, the model exhibited strong generalization capabilities when applied to an external 3D dataset, confirming its reusability beyond plant phenotyping tasks. Integrating this simple method into phenotyping pipelines will enhance efficiency and accuracy in high-throughput trait estimation, supporting advancements in plant science and precision agriculture.

INDEX TERMS Background segmentation, 3D imaging, Point cloud processing, Multi-Layer perceptron, Machine learning, Plant phenotyping, Remote sensing, Precision agriculture, Artificial intelligence

I. INTRODUCTION

Phenomics, particularly plant phenotyping, is an emerging research discipline focusing on understanding plant-based systems' dynamics. It is frequently aided by imaging technologies that generate a lot of data, which, in turn, requires processing using fit-for-purpose computational methods. The phenomics tasks support a range of applications in plant-related research disciplines, particularly those requiring high throughput and non-destructive plant monitoring – e.g., crop characterization for breeding, genebanks, agronomy, or basic research of plant-based system dynamics (e.g., stress responses to biotic/abiotic stimuli). Much of such research is done with sensors that capture 2D/3D RGB/multi-/hyper-spectral reflection of the plant systems (i.e., digital twins). While 2D imaging has been intensively explored for decades [1], the utilization of information from 3D imaging methods (e.g., LiDAR) is lagging behind, particularly for the new generation of sensors that also capture color and other information (e.g., near-infrared). In this paper, we particularly focus on the 3D plant scans. Those scans usually contain information on the background surrounding the plant, such as soil, a growing container, irrigation tubes, etc. Therefore, this background must be identified (segmented) and removed for further analysis to assess the plant characters per se.

Popular methods used to gather 3D data reconstruction are based on multiple 2D images taken from different positions – see categorization by [2]. Those systems capture 2D images of a plant from various angles and process them to create 3D models. The photogrammetry methods usually provide high-quality models. On the other hand, these systems do not

provide high throughput. This paper primarily focuses on laser-based scanner data originating from high-throughput scanners like Phenospex's (PSX) PlantEye(R) used for plant phenotyping. These types of scanners are usually mounted on stationary gantries with fixed positions and used for more accurate evaluation of crops.

Extensive research is being conducted in automotive and robotics, using mainly LiDAR (Light Detection and Ranging) sensors. In this field, various methods, such as filtering, have been used to preprocess the raw point cloud data [3]. Ground segmentation in 3D point clouds is also addressed throughout the literature. Most of the methods were developed for the automotive LiDAR datasets, e.g., [4], [5], [6], [7], [8], [9], [10]). Other authors address the area of robotics, e.g., [11],[12]. In the mentioned studies, LiDAR mainly captures information about moving scenes where the scanner and/or objects are not static. Thus, the analytics suitable for LiDAR image analysis might not fit all 3D data types. The abovementioned methods consider the ground as a flat surface that is narrow or tilted. In the plant-systems analysis (e.g., phenotyping), non-flat soil is often surrounded by additional background data. In plant-system analysis, LiDAR is frequently mounted on aerial moving vehicles like UAVs, tractors, or planes. It is used to infer plant canopy traits like plant height or canopy cover over large areas.

Various methods for background segmentation are suitable for different types of data – in general, these can be methods based on thresholding (i.e., height, the color of the individual points) or methods based on the image features that might be evolving/considering combinations of point-cloud features and their hidden patterns. In plant-systems analyses, authors mostly use direct methods to segment/detect whole plants, their organs, or roots, e.g., [13], [14], [15], [16], [17]. In our previous studies [18], [19], we utilized Region Growing Segmentation (RGS) [20] and Random Sample Consensus (RANSAC) [21], which are widely recognized and frequently applied algorithms for background segmentation in the literature. However, to achieve satisfactory results with these methods, the structure of our dataset required the separation of growing containers based on their mathematical coordinate information prior to segmenting background and plant data. Additionally, it was observed that these methods are not well-suited for scenarios involving small plants or inclined/wavy soil surfaces.

A typical simplistic segmentation strategy, which is supported by most preprocessing pipelines, such as Phenospex's Phena, is to divide plant and soil by employing a specified cutoff height (z-coordinate). The upper portion considered plant data, and the lower portion as background. The cut height is usually set above the rim of the tray to get rid of the tray points fully. However, as shown in Figure 1, in certain situations (e.g., early growth stages, prostrate type of plant canopy architectures), part of the plant is below the rim of the tray. This cut can lead to data loss. Therefore, this approach often leads to inaccurate data analyses and trait estimation. Additionally, the optimal separation height must be determined for each setup, complicating standardization and requiring precise repetition across experiments.

FIGURE 1. A scan of a small plant that grows below the rim of the tray. Traditional coordinate-based soil segmentation would cut the scan above the rim and lead to significant plant data loss.

In summary, existing algorithms are either very complex, developed for naturally different data, data lossy, or too simple that they need custom preprocessing setup for each different experiment type, i.e., setup of different coordinates for different growing containers, positioning, etc. The presented paper focuses on plant scans in 3D point clouds obtained by high-throughput phenotyping platforms. In particular, we use data from Phenospex's PlanEye F600 scanning technology [22] and the installation in Hyderabad, India [23]. While the platform was validated and is being deployed for a range of end-uses [24], [25], [26], some of the uses are currently constrained by the accuracy of the plant feature inference algorithms. As illustrated in FIGURE 1, this static method causes substantial data loss, leading to inaccuracies in the plant trait assessment.

The presented research aims to address two current gaps – an accuracy for a wide range of data while maintaining the necessary performance for high-throughput systems. The main objective of this paper is to provide an AI-driven method for segmenting the background from the plant data points for a wide range of species. This method can be used during data preprocessing before other analytical algorithms provided by phenotyping platforms. Moreover, the proposed method addresses the limitations of traditional lossy threshold-based segmentation methods. Additionally, we evaluate the efficiency in several ways, i.e., AI model evaluation using a test set, point counts comparison to the coordinate-based method, canopy trait inference using destructive measurement, and finally, a generalization capacity using an external, different-domain data set.

II. MATERIALS AND METHODS

This section describes the methodology used for background segmentation in 3D plant point clouds. It covers data acquisition using a high-throughput scanning platform, preprocessing steps to rotate the raw scans, manual annotation for ground truth generation, dataset partitioning for model training, and model performance and accuracy evaluation. The entire workflow is summarized in FIGURE 2.

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3594406

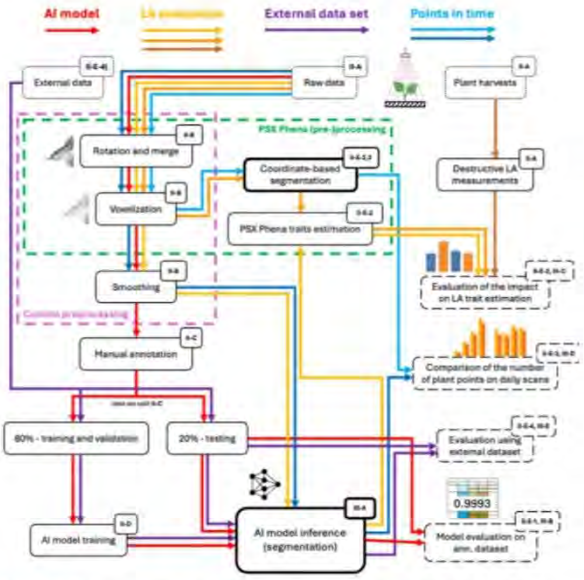


FIGURE 2. Visualization of the whole methodology. It shows individual steps (referenced by section numbers) and paths for different data visualized by colors. The colors, together with the source data, are referred to in TABLE 1.

A. DATA COLLECTION

The data used in this study was generated using a LeasyScan platform (ICRISAT, Hyderabad, India; Facilities & Services – GEMS, details in [23]), using PlantEye F600, a 3D scanner with multispectral imaging (PlantEye F600 multispectral 3D scanner for plants - PHENOSPEX). LeasyScan uses a dual scanning system (2 partially overlapping scanners capture the same area, mounted on the construction shown in FIGURE 3) to capture the 3D reflection of the target area and the capacity to scan an area of ~2500m^2 in 1h30min. As shown in FIGURE 3, the scanned area was equipped with the containers in which the crops were raised. In our study, we used a setup of microplots, each consisting of 40x60x60 cm blue plastic trays (blue containers in FIGURE 3) with ~ 70 kg of vertisol equipped with drip irrigation tubes. The scanning area was configured to capture the area delineated by individual microplots up to 1.5m from the ground via HortControl software. The area is divided into sectors that contain a certain number of microplots (24 divided into two rows for this study). The sectors are marked by physical barcodes that are recognized by the scanners. Finally, each scan consists of two separate *.ply files angled towards each other that cover one sector.

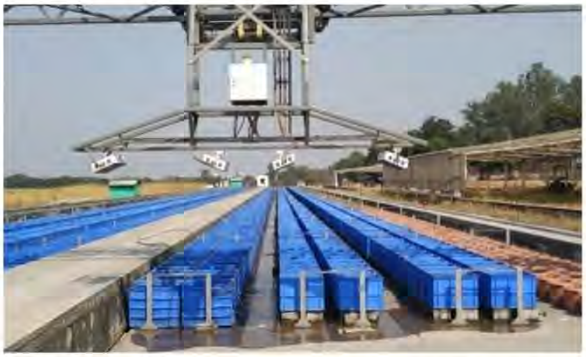


FIGURE 3. The LeasyScan high-throughput phenotyping platform was used to gather the data. The picture shows the dual position (twice two complementary, partially overlapping scanners capture the same area). The mounted scanners are moving over the field to capture the data (~2 500m2 area in 90 minutes).

In this study, we used scans from 4 crop species with different growth habits and canopy architectures (two cereals – pearl millet and sorghum, and two legumes – chickpea and mungbean). The crops were sown throughout March-April 2022 and raised with the standard fertilizer inputs under irrigated conditions (every 2-3 days). Each crop was sown to multiple microplots. Altogether, we used scans from 6 different experiments to train the background segmentation algorithms and evaluate plant canopies inference from 3D point clouds where the canopies' total leaf area (LA) from multiple plants in each microplot was measured destructively (Table 1). Each sector's total leaf area variability was created for the latter by i) different crop densities (2-6 plants/microplot) and ii) sequential crop harvests at four different dates. At these dates, the total crop in each microplot was harvested, and the leaf area was measured destructively by Li-Cor 4100. Altogether, we gathered 275 destructive LA measurements.

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3594406

TABLE 1. Overview of the unique experiments, sectors (identified by their barcode numbers), and a number of scanned microplots used throughout the paper for model training and evaluations. The colored arrows correspond to the arrows in FIGURE 2

Used for	Specie	Experiment IDs	Barcode IDs	Number of microplots
AI model training →	Pearl millet	50,51	28,96	96
	Sorghum	48	84	48
	Chickpea	57	83, 84, 86	216
	Mungbean	59	79 - 92	336
Leaf Area trait evaluation (destructive measurement) →→	Pearl millet	56	178, 179, 180	60
	Sorghum	56	154, 176, 177	72
	Chickpea	56	125, 126, 127	72
	Mungbean	56	151, 152, 153	72
Point counts on daily scans →→	Chickpea	57	80	24

B. DATA PREPROCESSING

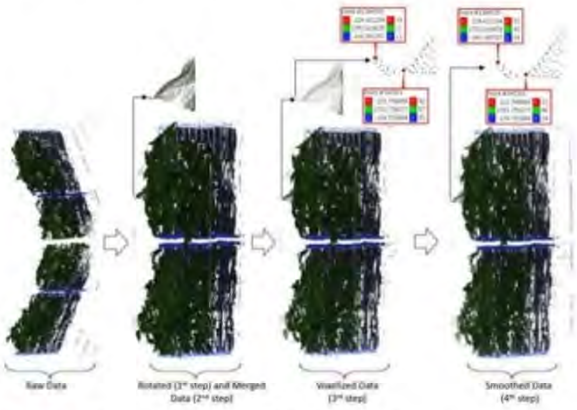


FIGURE 4. Preprocessing steps of the raw scan files to extract only plant data for individual microplots.

We used the raw scans obtained from the scanners. The whole preprocessing is visualized in FIGURE 4. Here, the first step involves rotating the scans to align flatly on the x-plane. Both scans are merged into a single file in the second step. This merging process increases the point cloud density in the overlapping areas scanned by both scanners. Therefore, the third step involves a voxelization process (dividing the 3D space into small cubes called voxels, each producing a point cloud value representing all points within it) to rearrange the points in space uniformly. The efficiency of the voxelization has been proven by, e.g., [27], [28], [29]. In this study, for instance, voxelization reduced the number of points in a single scan from 16,735,700 to 2,677,885. This step is performed by a bespoke version of Phenospex's Phena v2 pipeline, which the vendor customized for us to ensure backward compatibility with their systems.

After the scanning process, specific points may be considered outlier values, where the color values differ significantly from the others. Since the developed AI model uses color values, plant, soil, and tray color values must be consistent within themselves. For this purpose, a smoothing process (4^{th} step) was applied to eliminate outlier color values in some points. In this process, each point takes the average color value of the n nearest point. In the cropped voxelized leaf section shown in FIGURE 4, there are approximately 1000 points. Based on this density, the value of N has been set to 250 in this study. In the zoomed-in view of the cropped leaf area in FIGURE 4, the color values of the points before and after the smoothing process are given, demonstrating how the smoothing process ensures that the points on the same leaf have similar color values. This approach is consistent with previous studies, where smoothing techniques have been utilized to reduce noise and irregularities in point cloud color values, improving both visual consistency and feature reliability [30], [31].

C. DATA ANNOTATION AND SPLIT

The plant and background regions were manually marked and separated using Cloud Compare (version 2.10). The annotated sector is depicted in FIGURE 5. This process involved precisely defining the plant's boundaries and background data. Regions identified as plant data were labeled and stored separately from the background data. In instances where certain regions could not be distinguished, these ambiguous areas were excluded from the dataset to avoid model confusion. This process was performed for each plant species and all sectors. In total, we used 696 microplots, as depicted in TABLE 1.

FIGURE 5. The top image (a) shows the manually separated plant data, while the bottom image (b) shows the background data.

Following the annotation process, the dataset was randomly split into three subsets for use in the AI model: 60% for training, 20% for validation, and 20% for testing. These scans, detailed in Table 1, were selected based on the following criteria to ensure a balanced and representative dataset. Since

there was an unequal number of available scans for each plant species, fewer files were included for some species, such as sorghum and pearl millet. Additionally, given the differing growth rates of each plant species, scans were typically chosen a few days before harvest, as this period often provides the most representative plant structure for segmentation tasks. For species with fewer available scans, additional files were selected from earlier time points within the same experiment to enhance diversity and representation. These selections were made by considering factors such as plant species, the number of available scans, and growth rates, ensuring a dataset that reflects the variability of plant development.

D. AI MODEL DEVELOPMENT

The experiments were conducted using artificial neural networks, specifically a Multi-Layer Perceptron (MLP) model. Keras hyperparameter tuning was configured using Bayesian Optimization, with a maximum of 500 trials to search the hyperparameter space efficiently. The hyperparameters considered are presented in TABLE 2. Due to the binary classification task, the output layer consisted of a single neuron responsible for the final prediction (plant/background). The maximum number of epochs was set to 50, and an early stopping method was employed to prevent overfitting and improve generalization. The early stopping was based on validation loss, with a patience period of 2 epochs. The model training was performed separately for three input point data: RGB, RGB+XYZ, and RGB+XYZ+NIR (where RGB represents color channels, XYZ corresponds to 3D positional coordinates, and NIR indicates near-infrared value).

TABLE 2. Range of the hyperparameters used by Keras tuner during the model development.

Hyperparameter	Range of Possible Values (well known)
Number of hidden layers	1 to 3
Neurons per hidden layer	[10, 20, 50]
Activation function	Relu, Sigmoid, Tanh
Optimizer	SGD, RMSProp, ADAM
Input data	RGB, RGB XYZ, RGB NIR XYZ

E. MODEL EVALUATION

The performance of the developed model was evaluated from three perspectives. First, the performance of the MLP model using the annotated point clouds as ground truth; second, the impact of the segmentation on Leaf Area trait estimation using ground truth gathered by the destructive measurement method; third, we provide a comparison using external data set.

1) MODEL EVALUATION ON ANNOTATED DATASET

For the first case – the performance of the MLP model, we used mainly two well-known metrics: Accuracy (number of correct predictions divided by the total number of predictions) and Confusion matrix (a table displaying the ground truth versus predicted classifications). We also provide Precision and Recall calculated from the latter one. The evaluation used the test set defined in II-C Data annotation and split. We also evaluate the impact of the smoothing process by comparing it to the model trained without this preprocessing step.

2) EVALUATION OF THE IMPACT ON LEAF AREA TRAIT ESTIMATION

For the second case, we used the data that were scanned and harvested for the destructive measurements (ground truth). The data origins are visualized using orange arrows in TABLE 1 and FIGURE 2. In FIGURE 2 there are three paths. The darker one is for the ground truth. Light one as follows: 1) we ran the preprocessing; 2) we ran the model inference for each datapoint to classify it to separate plant and background data (shown at the top of FIGURE 4); 3) the plant data were then sliced up based on the fixed coordinates to get individual microplot data, as illustrated on the left side of FIGURE 6; 4) we predicted the Leaf Area trait using the customized version of Phena provided by Phenospex. For the middle orange, the whole Phena pipeline was used, utilizing the coordinate-based segmentation method. However, as we used the version of the Phena pipeline that is not yet in production, we preprocessed the scans using voxel resolutions from 0.55 to 0.65 in both paths.

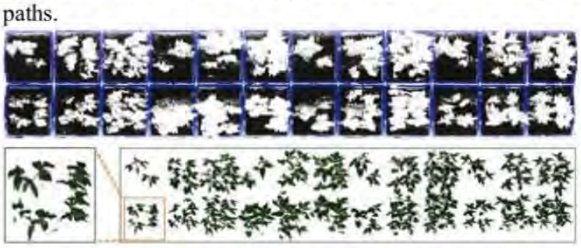


FIGURE 6. Separated background and plant data. On the left side, the zoomed-in points show the plant data belonging to a single microplot separated from the full sector scan on the right side. Similarly, the data of all microplots are separated from each other and saved as a separate file for evaluation.

Finally, we compared the leaf area trait measured by the destructive method on each microplot to the estimations computed by the Phena pipeline (using coordinate-based segmentation) and our proposed method (AI-based segmentation). The comparison was evaluated using standard, well-known metrics, particularly R^2 and root mean square error (RMSE). These metrics were calculated for all the voxel resolutions (0.55 to 0.65). Because of the inherent differences between the canopy structures of different crop species and related bias in the destructive measurement of the ground truth, we provide evaluation separately for cereals (pearl millet, sorghum), broad-leaf legumes (mungbean) and narrow-leaf legume (chickpea).

3) COMPARISON OF THE NUMBER OF PLANT POINTS ON DAILY SCANS

The trained AI model was additionally tested on everyday scans throughout the growing period using a previously unseen chickpea setup (Experiment 57, barcode No. 80). The segmentation was performed using both methods. The scan

files were segmented using both methods from the initial sowing day to harvest. The point cloud counts of the separated plant data were recorded for each day. For the coordinate-based method, the segmentation threshold was manually tuned and set at the most precise z-coordinate of 118, corresponding to the microplot's optimal segmentation height.

4) EVALUATION USING EXTERNAL DATASET

To show the generalization capacity and effectiveness of the proposed model, we provide an additional evaluation using a dataset from a different domain. The model was re-trained using the 3D Paris-Lille dataset collected by [32] from the streets of Paris and Lille. This dataset encompasses three point cloud files: Lille1.ply, Lille2.ply, Paris.ply, and 50 distinct classes. The original dataset publication did not provide a baseline classification result; however, in subsequent studies, the results obtained with this dataset have been assessed in different ways. For instance, in the study by Diaz et al. (2021), they trained their model on a dataset they created themselves and used the 3D Paris-Lille dataset solely to evaluate the model's performance. While Diaz et al., in the original study, specify that they used 7 data files for training and 3D Paris-Lille data files for testing, these training datasets were not available in their shared GitHub repository. Therefore, we divided the dataset into 80% for training and validation and 20% for testing to align with our study's needs. In our case, we trained our model from scratch using this training data.

This approach allowed us to proceed with model evaluation but complicated a direct comparison with the accuracy values reported in their study. While applying the pipeline to this dataset, we only modified the input data to align with the dataset's characteristics. Then, the dataset was processed. Considering these limitations, we evaluated the model's performance by examining the average accuracy instead of directly comparing our results with those of [33].

III. RESULTS

A. FINAL MODEL

The optimal model configuration consists of three hidden layers: 10 units in the first layer and 50 units in both the second and third layers, all using the ReLU activation function. The output layer used the sigmoid activation function, which is suitable for binary classification. The model was trained using the Adam optimizer with binary cross-entropy as the loss function. This configuration, derived from the RGB+XYZ+NIR dataset, demonstrated the best performance. The model's architecture is visualized in FIGURE 7. The final trained model and the source code (preprocessing and model) can be found in a public GitHub repository (https://github.com/serkankartal/MLP-3D-PlantSeg). The repository also contains instructions on how to train the model for custom data.

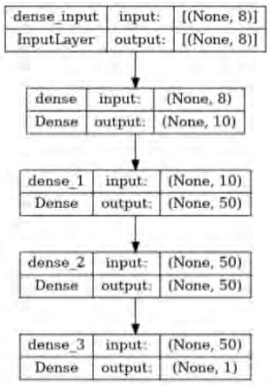


FIGURE 7. The architecture of the MLP model used for background segmentation. The input consists of eight neurons representing three RGB channels, one NIR channel, and spatial information (XYZ). After three hidden layers, one output layer represents the final category – background/plant.

Optimizing the model with Keras Tuner (for 500 trials) on an NVIDIA A4500 GPU took approximately 10 hours for each dataset. For the soil segmentation task, the trained model processes one point at a time to classify it as plant or background data. As the model is relatively simple, the execution time for one-point prediction is usually in milliseconds, based on the hardware used. Inference of a single scan file (whole sector), including preprocessing, takes about 2 minutes on average, with most time spent on preprocessing rather than the AI model.

B. MODEL EVALUATION ON ANNOTATED DATASET

Two widely recognized metrics were used to evaluate the model's performance on the annotated dataset: Confusion Matrix and Accuracy. The assessment used a test set comprising 20% of the manually annotated scans. The best model configurations identified during the hyperparameter tuning process were evaluated. TABLE 3 shows a comparison of the models using different input point data.

TABLE 3. Confusion matrices and accuracies to compare various variants of models – input data.

M		Ground Truth		
Input point data	Accuracy	0	1	Model
Unsmoothed RGB + XYZ + NIR	0.9899	18,202,033	169 835	0
		98,471	8,072,451	1
RGB	0.9958	18,233,465	44,276	0
		67,039	8,198,010	1
RGB + NIR	0.9970	18,247,405	41,614	0
		50,646	8,203,125	1
RGB + XYZ	0.9989	18,284,867	14,990	0
		13,828	8,229,105	1
RGB + XYZ + NIR	0.9993	18,296,710	11,159	0
		7,005	8,227,916	1

All the models consistently demonstrated high classification accuracy, with performance exceeding 99% in each case. Although the differences in accuracy between the different inputs were minimal, a closer look at the confusion matrices reveals that the spatial point data (XYZ) and near-infrared spectra (NIR) provided significant information to the model. For example, the false positives decreased significantly from 67,039 in the RGB model to 7,005 in the RGB+XYZ+NIR model, while false negatives dropped from 44,276 to 11,159. This additional input point data improved accuracy and enhanced precision by reducing misclassifications. Therefore, the model trained on the RGB+XYZ+NIR dataset was selected for further use in this study.

TABLE 3 also shows the specific impact of the smoothing preprocessing step. We evaluated the best model trained on unsmoothed RGB + XYZ + NIR input point data. This model produced 98,471 false positives and 169,835 false negatives. The accuracy dropped to 0.9899. This sharp contrast in performance indicates that the lack of smoothing introduced substantial noise into the classification process, leading to a much higher rate of misclassifications. Thus, the smoothing process reduced outlier effects by averaging the color values of neighboring points.

C. EVALUATION OF THE IMPACT ON LEAF AREA TRAIT ESTIMATION

For the second evaluation, we used the destructively measured dataset, as defined in TABLE 3, to compare the estimation of the Leaf Area trait. FIGURE 8 presents the resulting metrics (MAPE and R^2) of selected voxel resolution values. We picked 0.60 as the preprocessing value for each species for the model development, the best value for our AI-based method, and the best value for the coordinate-based method. Results for all species and resolutions can be seen in the supplementary file LA_results.xlsx.

When the results for Mungbean, Pearl millet, and sorghum are examined, it is observed that the error rates in the leaf area indices obtained using the AI-based method we proposed and the traditional method are close to each other. The main reason is that these three species relatively quickly grow above the tray rim. This fact significantly affects the results. Nevertheless, the predictions made with the AI-based method still provide better results (except for mungbean).

However, when the metric values obtained for the chickpea are examined, it is seen that the AI-based method shows higher differences. For example, at a resolution of 0.62, the R2 value for the AI-based method is 11.35% higher than the best value for the Coordinate-based method at 0.65. This difference demonstrates an improvement in prediction accuracy for the AI-based method, especially when dealing with smaller plants like chickpeas.

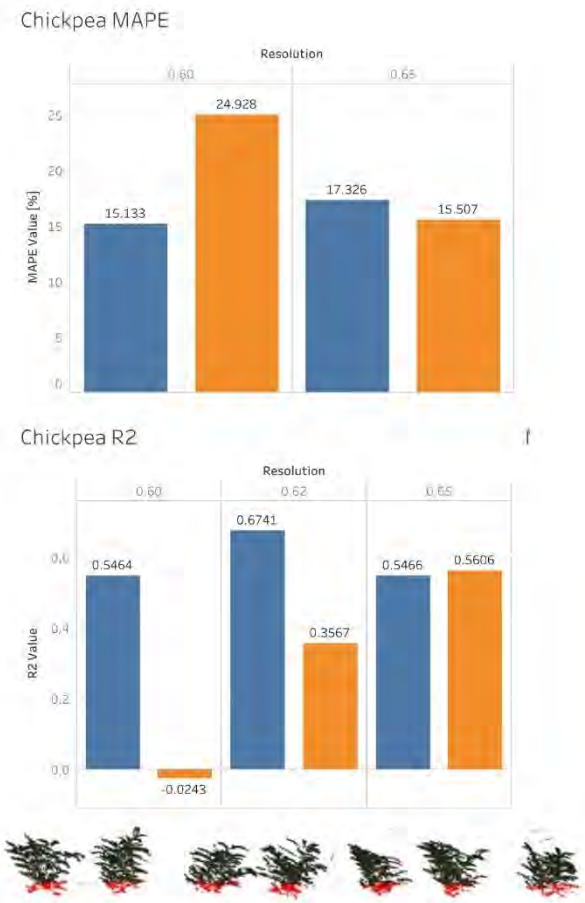


FIGURE 8. Results of Leaf Area trait evaluation using mean absolute percentage error (MAPE) and R2. We show only specific resolution values, particularly 0.60 used in AI-based method development), the best value for model inference, and the best value for the coordinate-based method. Below, there is a visualization of the difference using both methods for segmenting plants. The plants were scanned on the 22nd day from sowing. The red part represents points below the tray rim that are not segmented by the coordinate-based method.

A visual comparison of the segmented points is also provided in FIGURE 8. As seen in the figure, the points highlighted in red at the bottom of the image represent the additional plant data selected by the AI-based method, i.e., the points that could be segmented from within the tray. This visualization clearly demonstrates the source of the extra plant data obtained by our proposed AI-based method.

D. COMPARISON OF THE NUMBER OF PLANT POINTS ON DAILY SCANS

The comparison of the point counts is detailed in FIGURE 9. In the early growth stages, when chickpea plants emerged and remained confined within the microplot, the coordinate-based method inaccurately classified barcode and noise data above the microplot as plant data, leading to a count of 6,228 points instead of 0. On Day 9, the AI-based method identified 1,100,423 plant points, while the coordinate-based method

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3594406

identified 994,136 points, resulting in a 10.7% increase in plant data segmentation. On Day 11, this difference was even more pronounced, with the AI-based method identifying 846,792 points compared to 719,871 points segmented by the coordinate-based method, a 17.6% increase. Similarly, on Day 20, the AI-based method segmented 2,463,856 points compared to 2,367,033 points by the coordinate-based method, representing a 4.1% improvement. A decrease in the number of segmented points was caused by the thinning process performed on Days 10 and 11, which removed excess plants.

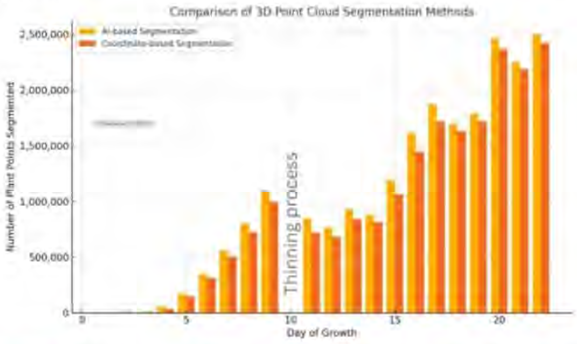


FIGURE 9. Comparison of both segmentation methods in daily plant point cloud counts using independent chickpea experiment. On day 10, the decrease in count is caused by the manual thinning process, which removes excess plants. The reason for the difference is illustrated in FIGURE .

The coordinate-based method in the early growth stages, when plants have not yet grown above the tray rim, captures noisy data, such as that of irrigation pipes. Thus, the number of points does not differ from the AI-based method or is even higher for the coordinate-based method. The AI-based method, on the other side, accurately segmented only the plant data inside the microplot. This highlights the AI-based method's capability to distinguish plant data from background and noise without relying on thresholds or manual adjustments. Such precision is particularly crucial for early-stage plant monitoring, where the traditional methods often fail to detect significant portions of plant data within the microplot.

E. EVALUATION USING EXTERNAL DATASET

The results obtained and reported on our pipeline are derived exclusively from the test dataset (as defined in section II-E-2)). The proposed pipeline achieved an average accuracy of 0.98, which is similar to the 0.97 accuracy reported for the entire dataset. Additionally, the result of the separation performed by the model on the Paris Street data is illustrated in FIGURE 10. In this figure, the blue regions represent background data, while the red areas denote other data (trees, cars, streetlights, etc.). The visual analysis effectively showcases the model's success in separating the ground from non-ground elements, highlighting the robustness and versatility of the AI-based pipeline across various datasets.

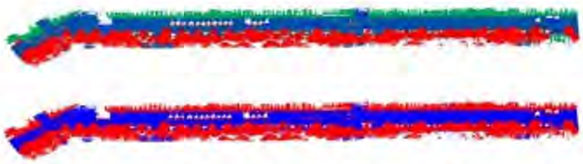


FIGURE 10. The visualization at the top shows the classified data of the original dataset. In contrast, the image at the bottom shows the dataset processed by our model into two categories (ground and non-ground) according to the format suitable for our study.

IV. DISCUSSION

A. SEGMENTATION FOR PLANT PHENOMIC TASKS

The findings of this study highlight key challenges in 3D plant phenotyping related to background segmentation in point cloud data obtained from high-throughput laser-based scanners. Traditional segmentation methods rely on height coordinates. Those methods often result in significant data loss, particularly for early-stage plants or prostrate canopy architectures. Additionally, commonly used algorithms like Region Growing Segmentation (RGS) and Random Sample Consensus (RANSAC) may not generalize well across different experimental setups. Computer vision-based methods, e.g., [13], [14], [15], [16] are very complex and rely on labor-intensive annotation. Existing solutions in automotive and robotics applications [4], [5], [6], [7], [8], [9], [10], [11], [12] focus on flat surfaces and dynamic environments, making them less suitable for plant phenotyping, where soil surfaces are often uneven, and additional background noise is present.

B. SIGNIFICANCE OF THE RESULTS

This study presents an AI-driven segmentation method designed to improve the accuracy of traditional coordinate-based segmentation methods while addressing the abovementioned issues. The proposed method has been rigorously evaluated through AI model performance metrics, comparison with coordinate-based segmentation, canopy trait inference, and external dataset validation, demonstrating its potential to enhance plant feature extraction in phenotyping workflows.

We have shown the results of a simple MLP-based model. The RGB+XYZ+NIR input point data configuration achieved a classification accuracy of 0.9993, with only 7,005 false positives (points) and 11,159 false negatives, significantly outperforming more straightforward configurations. This result highlights the importance of integrating spatial and spectral data to enhance model precision and reliability. Besides the segmentation model, we also implemented a preprocessing step smoothing to reduce noise and irregularities in point cloud color values. The smoothing step improved the model's accuracy by 0.0094%, reducing 87,312 false positives and 158,676 false negatives.

The method's impact on the accuracy of crops' leaf area estimation was evident. The accuracy of the algorithms presented for leaf area estimation in crops such as pearl millet

and mungbean was comparable to that of the traditional coordinate-based method. This is because these crop seedlings grow rapidly in a vertical direction above the tray rim into the space where the canopy is captured by coordinate-based algorithms. In smaller crop species with prostrate growth habits, such as chickpeas, there were significant improvements in accuracy in capturing the canopy area. For chickpeas, the hereby proposed method achieved an R^2 value 0.1135 higher than the coordinate-based approach and segmented 7% more plant data that were below the tray rim. Such magnitude of differences in canopy size for small and prostrate crops like chickpeas is important for crops' capacity to adapt to abiotic stresses like drought [26], [34]. This indicates the hereby presented method significantly improves crop evaluation PSX's platforms, such as LeasyScan, particularly the ones with small and prostrate canopy structures. This was demonstrated in the case of chickpeas. Finally, we expect that the accuracy gains can be even higher for crops like peanuts or wild crop relatives.

C. SOURCE CODE AND HOW TO USE THE MODEL

The hereby reported algorithm is available on the GitHub repository (https://github.com/serkankartal/MLP-3D-PlantSeg). The model is simple to reuse, re-train, or fine-tune for similar types of data. Instructions are in the repository README file, including the specific information for users with data from Phenospex's PlantEye scanner. We aim for a dynamic, continuous, feedback-based future development. Any feedback is welcomed by e-mail to the corresponding author or preferably via Issues on the GitHub platform, where everyone can see answers.

D. GENERALIZATION CAPACITY

The model also demonstrated its generalization capacity when tested on an external source of 3D point cloud data – the Paris-Lille 3D dataset by [32], achieving an accuracy of 0.98, which is similar to the 0.97 reported by [33]. This evaluation needed model re-training as the external dataset represents a different and broader range of data (trees, streets, sidewalks, traffic signs, lightning, etc.). Despite the re-training need, the results confirm the robustness and adaptability of the proposed model to diverse datasets, making it suitable for applications beyond agriculture/phenomics. The model, in general, can be used for any 3D point cloud data.

E. LIMITATIONS AND FURTHER IMPROVEMENTS

While the proposed method shows substantial accuracy (see section C) and reasonable performance, certain limitations remain. Although we selected species to cover a wide range of crop canopy types (grain legumes and cereals), the model was trained and evaluated on a limited number of species. Additional improvements to robustify and generalize the model, especially for more species, canopy types, and different types of trays, can be performed to transfer the method into the production-ready phase. Specifically, further resolution (voxel sizes) optimization for specific (small-sized, prostrate) crop types or growth stages could enhance accuracy. Using additional training data with more diverse crop types (e.g., peanut or wild crop progenitors) might improve the model's generalization capacity.

V. CONCLUSION

This study presents a novel AI-driven (MLP-based) method for background segmentation in 3D plant phenotyping that addresses the limitations of traditional coordinate-based methods. Using a Multi-Layer Perceptron with integrated RGB, spatial (XYZ), and near-infrared (NIR) inputs, our approach achieves a classification accuracy of 0.9993.

We innovatively used the smoothing process in the preprocessing stage, effectively reducing noise by averaging color values across neighboring points (accuracy increased by 0.0094). This enhancement, combined with integrating spatial and spectral data, leads to a more reliable separation of plant and background regions, particularly for small or prostrate species where a particular part of the plant is below the growing tray.

The improved segmentation accuracy directly benefits trait estimation; for example, our method increased the R^2 value by 0.114% in chickpea crops and captured an additional 7.01% of plant data compared to the traditional coordinate-based method. We also showed the model's ability to adapt to domain shift by re-training using the external dataset, confirming its generalization capacity.

Overall, this paper offers an efficient novel method usable for automating plant trait analysis in high-throughput phenotyping. The complete model and code are available in our public GitHub repository for possible feedback and future development.

ACKNOWLEDGMENT

The authors thank Phenospex (the manufacturer of the PlantEye 3D scanners), mainly Thorsten Karrer, Christiaan Vonk, and András Tóth. for providing a customized version of the Phena pipeline that allowed the Leaf area trait evaluation.

REFERENCES

[1] T. Meraj, M. I. Sharif, M. Raza, A. Alabrah, S. Kadry, and A. H. Gandomi, "Computer vision-based plants phenotyping: A comprehensive survey," *iScience*, vol. 27, no. 1, p. 108709, Jan. 2024, doi: 10.1016/j.isci.2023.108709.

[2] F. Okura, "3D modeling and reconstruction of plants and trees: A cross-cutting review across computer graphics, vision, and plant phenotyping," *Breed Sci*, vol. 72, no. 1, p. 21074, 2022, doi: 10.1270/jsbbs.21074.

[3] X.-F. Han, J. S. Jin, M.-J. Wang, W. Jiang, L. Gao, and L. Xiao, "A review of algorithms for filtering the 3D point cloud," *Signal Process Image Commun*, vol. 57, pp. 103–112, Sep. 2017, doi: 10.1016/j.image.2017.05.009.

[4] J. Cheng, D. He, and C. Lee, " A simple ground segmentation method for LiDAR 3D point clouds," in *Proceedings - 2020 2nd International Conference on Advances in Computer Technology, Information Science and Communications (CTISC 2020)*, 2020, pp. 171–175.

[5] S. Choi, J. Park, J. Byun, and W. Yu, "Robust ground plane detection from 3D point clouds," in *2014 14th International Conference on Control, Automation and Systems (ICCAS 2014)*, IEEE, Oct. 2014, pp. 1076–1081. doi: 10.1109/ICCAS.2014.6987936.

[6] W. Huang *et al.*, "A Fast Point Cloud Ground Segmentation Approach Based on Coarse-To-Fine Markov Random Field," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 7841–7854, Jul. 2022, doi: 10.1109/TITS.2021.3073151.

[7] K. Liu, W. Wang, R. Tharmarasa, J. Wang, and Y. Zuo, "Ground Surface Filtering of 3D Point Clouds Based on Hybrid Regression Technique," *IEEE Access*, vol. 7, pp. 23270–23284, 2019, doi: 10.1109/ACCESS.2019.2899674.

[8] P. Narksri, E. Takeuchi, Y. Ninomiya, Y. Morales, N. Akai, and N. Kawaguchi, "A Slope-robust Cascaded Ground Segmentation in 3D Point Cloud for Autonomous Vehicles," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, IEEE, Nov. 2018, pp. 497–504. doi: 10.1109/ITSC.2018.8569534.

[9] Y. Qian, X. Wang, Z. Chen, C. Wang, and M. Yang, "Hy-Seg: A Hybrid Method for Ground Segmentation Using Point Clouds," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 2, pp. 1597–1606, Feb. 2023, doi: 10.1109/TIV.2022.3187008.

[10] Z. Shen, H. Liang, L. Lin, Z. Wang, W. Huang, and J. Yu, "Fast Ground Segmentation for 3D LiDAR Point Cloud Based on Jump-Convolution-Process," *Remote Sens (Basel)*, vol. 13, no. 16, p. 3239, Aug. 2021, doi: 10.3390/rs13163239.

[11] H. Vu *et al.*, "Adaptive ground segmentation method for real-time mobile robot control," *Int J Adv Robot Syst*, vol. 14, no. 6, p. 172988141774813, Nov. 2017, doi: 10.1177/1729881417748135.

[12] H. Vu, H. T. Nguyen, P. Chu, S. Cho, and K. Cho, "A Ground Segmentation Method Based on Gradient Fields for 3D Point Clouds," 2018, pp. 388–393. doi: 10.1007/978-981-10-7605-3_64.

[13] K. Mirande, C. Godin, M. Tisserand, J. Charlaix, F. Besnard, and F. Hétroy-Wheeler, "A graph-based approach for simultaneous semantic and instance segmentation of plant 3D point clouds," *Front Plant Sci*, vol. 13, Nov. 2022, doi: 10.3389/fpls.2022.1012669.

[14] L. Wang, L. Zheng, and M. Wang, "3D Point Cloud Instance Segmentation of Lettuce Based on PartNet," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, Jun. 2022, pp. 1646–1654. doi: 10.1109/CVPRW56347.2022.00171.

[15] R. Zhang, Y. Wu, W. Jin, and X. Meng, "Deep-Learning-Based Point Cloud Semantic Segmentation: A Survey," *Electronics (Basel)*, vol. 12, no. 17, p. 3642, Aug. 2023, doi: 10.3390/electronics12173642.

[16] J. Zhou, X. Fu, S. Zhou, J. Zhou, H. Ye, and H. T. Nguyen, "Automated segmentation of soybean plants from 3D point cloud using machine learning," *Comput Electron Agric*, vol. 162, pp. 143–153, Jul. 2019, doi: 10.1016/j.compag.2019.04.014.

[17] K. Zarzyńska, D. Boguszewska-Mańkowska, and A. Nosalewicz, "Differences in size and architecture of the potato cultivars root system and their tolerance to drought stress," *https://pse.agriculturejournals.cz/doi/10.17221/4/2017-PSE.html*, vol. 63, no. 4, pp. 159–164, 2017, doi: 10.17221/4/2017-PSE.

[18] S. Kartal *et al.*, "Machine Learning-Based Plant Detection Algorithms to Automate Counting Tasks Using 3D Canopy Scans," *Sensors*, vol. 21, no. 23, p. 8022, Dec. 2021, doi: 10.3390/s21238022.

[19] S. Kartal, S. Choudhary, M. Stočes, P. Šimek, T. Vokoun, and V. Novák, "Segmentation of Bean-Plants Using Clustering Algorithms," *Agris on-line Papers in Economics and Informatics*, vol. 12, no. 3, pp. 36–43, Sep. 2020, doi: 10.7160/aol.2020.120304.

[20] L. Garcia Ugarriza, E. Saber, S. R. Vantaram, V. Amuso, M. Shaw, and R. Bhaskar, "Automatic Image Segmentation by Dynamic Region Growth and Multiresolution Merging," *IEEE Transactions on Image Processing*, vol. 18, no. 10, pp. 2275–2288, Oct. 2009, doi: 10.1109/TIP.2009.2025555.

[21] U. Weiss and P. Biber, "Plant detection and mapping for agricultural robots using a 3D LIDAR sensor," *Rob Auton Syst*, vol. 59, no. 5, pp. 265–273, May 2011, doi: 10.1016/j.robot.2011.02.011.

[22] Phenospex, " PlantEye F600 multispectral 3D scanner for plants - PHENOSPEX," https://phenospex.com/products/plant-phenotyping/planteye-f600-multispectral-3d-scanner-for-plants/.

[23] V. Vadez, J. Kholová, G. Hummel, U. Zhokhavets, S. K. Gupta, and C. T. Hash, "LeasyScan: a novel concept combining 3D imaging and lysimetry for high-throughput phenotyping of traits controlling plant water budget," *J Exp Bot*, vol. 66, no. 18, pp. 5581–5593, Sep. 2015, doi: 10.1093/jxb/erv251.

[24] V. Vadez, J. Kholová, G. Hummel, U. Zhokhavets, S. K. Gupta, and C. T. Hash, "LeasyScan: A novel

concept combining 3D imaging and lysimetry for high-throughput phenotyping of traits controlling plant water budget," *J Exp Bot*, 2015, doi: 10.1093/jxb/erv251.

[25] M. Tharanya *et al.*, "Quantitative trait loci (QTLs) for water use and crop production traits co-locate with major QTL for tolerance to water deficit in a fine-mapping population of pearl millet (Pennisetum glaucum L. R.Br.)," *Theoretical and Applied Genetics*, vol. 131, no. 7, pp. 1509–1529, Jul. 2018, doi: 10.1007/S00122-018-3094-6,.

[26] K. Sivasakthi *et al.*, "Plant vigour QTLs co-map with an earlier reported QTL hotspot for drought tolerance while water saving QTLs map in other regions of the chickpea genome," *BMC Plant Biol*, vol. 18, no. 1, 2018, doi: 10.1186/s12870-018-1245-1.

[27] H. Yang *et al.*, "Denoising of 3D MR Images Using a Voxel-Wise Hybrid Residual MLP-CNN Model to Improve Small Lesion Diagnostic Confidence," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13433 LNCS, pp. 292–302, 2022, doi: 10.1007/978-3-031-16437-8_28.

[28] Y. Liu *et al.*, "ET-PointPillars: improved PointPillars for 3D object detection based on optimized voxel downsampling," *Mach Vis Appl*, vol. 35, no. 3, pp. 1–13, May 2024, doi: 10.1007/S00138-024-01538-Y/METRICS.

[29] C. Lv, W. Lin, and B. Zhao, "Approximate intrinsic voxel structure for point cloud simplification," *IEEE Transactions on Image Processing*, vol. 30, 2021, doi: 10.1109/TIP.2021.3104174.

[30] I. C. Engin and N. H. Maerz, "Investigation on the processing of LiDAR point cloud data for particle size measurement of aggregates as an alternative to image analysis," *J Appl Remote Sens*, vol. 16, no. 01, Feb. 2022, doi: 10.1117/1.JRS.16.016511.

[31] H. Wang, F. Chen, W. Liu, and X. Zeng, "Unfolding Gradient Graph Regularization for Point Cloud Color Denoising," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* , vol. 15036 LNCS, pp. 565–579, 2025, doi: 10.1007/978-981-97-8508-7_39.

[32] X. Roynard, J.-E. Deschaud, and F. Goulette, "Paris-Lille-3D: A large and high-quality ground-truth urban point cloud dataset for automatic segmentation and classification," *Int J Rob Res*, vol. 37, no. 6, pp. 545–557, May 2018, doi: 10.1177/0278364918767506.

[33] N. Diaz, O. Gallo, J. Caceres, and H. Porras, "Real-time ground filtering algorithm of cloud points acquired using Terrestrial Laser Scanner (TLS)," *International Journal of Applied Earth Observation and Geoinformation*, vol. 105, p. 102629, Dec. 2021, doi: 10.1016/j.jag.2021.102629.

[34] M. Zaman-Allah, D. M. Jenkinson, and V. Vadez, "Chickpea genotypes contrasting for seed yield under terminal drought stress in the field differ for traits related to the control of water use," *Functional Plant Biology*, vol. 38, no. 4, 2011, doi: 10.1071/FP10244.

S. KARTAL Serkan Kartal received his BE (2010) and PhD (2017) degrees in Computer Engineering from Cukurova University, where he currently works as an associate professor in the Department of Computer Engineering. He conducted postdoctoral research at Delft University of Technology and the Czech University of Life Sciences in Prague, focusing on artificial intelligence applications in remote sensing and agriculture. His research interests include deep learning, computer vision, and AI-powered environmental monitoring. He is actively involved in several national and international R&D projects, including EU Horizon-funded initiatives in fisheries monitoring and sustainable agriculture.

J. MASNER Assoc. Prof. Jan Masner, Ph.D., is an expert in information management and modern information technologies, particularly UX and AI. He studied at the Czech University of Agriculture in Prague, where he successfully completed a doctoral program in Information Management. His extensive professional background, including participation in research projects funded by both Czech and international grant agencies and contractual research focused on digitization, user experience (UX), usability, and AI, while actively contributing to the development of innovative solutions in the digitization of public administration and precision agriculture, attests to his professional qualifications. In 2024, he obtained his habilitation in Systems Engineering and Informatics. Since then, he has been serving as an associate professor at the Department of Information Technologies, where he not only imparts his knowledge to students but also develops research activities with an emphasis on UX, usability, and AI. His research focuses on bridging information technologies with the life sciences. He deals in detail with computer vision, the integration of time series with computer vision methods, and the detection of plants and their organs from 3D scans, with a particular focus on innovative approaches to data augmentation.

J. KHOLOVA Dr. Jana Kholova received her MSc. (2006) and PhD (2010) in plant genetics and physiology from Charles University in Prague. In her current position at CZU and CATRIN, she coordinates trans-disciplinary, international research teams and integrate relevant tools and technologies to help communities address global challenges related to the sustainability of agricultural crop production systems in the face of climate change. She leads teams that support crop improvement programs, genebanks, agro-tech industries, primary crop producers, and basic research within academia, both in developing countries and globally.

A. GALBA RNDr. Alexander Galba is a graduate of the Faculty of Mathematics and Physics of Charles University in the field of theoretical cybernetics. At the Department of Information Technologies of the Faculty of Life Sciences of the Czech University of Life Sciences, he is engaged in research in the field of using computer vision in plant phenotyping. Primarily, he is developing advanced algorithms for processing 3D data for use in the field of AI. He is completing his doctoral studies in the program of systems engineering and informatics. His professional focus is on the area of connecting modern information technologies with practice in agriculture. He participates in projects of a data platform for storing and sharing research data, creating algorithms for modern 3D augmentation methods and using neural networks to solve phenotyping problems. He applies his research experience in teaching, where he focuses on foreign students and teaching English in subjects on the topic of information systems and programming. In addition to academic activities, he is also involved in commercial projects in the fields of information systems design, infrastructure, and applications. He holds certifications from HPE and Microsoft. He received a Microsoft Silver Award for his web application reservation system project.

T. MURUGESAN Dr. Tharanya Murugesan is a crop physiologist with over a decade of research experience, currently serving as an Associate Scientist at ICRISAT, India. She holds a Ph.D. in Biotechnology from Bharathidasan University, India, with her doctoral work focused on drought adaptation in pearl millet through integrated physiological, molecular, and genetic approaches. Her research expertise spans high-throughput phenotyping of cereal germplasm (pearl millet, sorghum, and foxtail millet), functional dissection of drought-tolerance QTLs, and mechanistic studies of plant water transport pathways. She has also pioneered the use of computer tomography imaging for rapid, non-destructive post-harvest phenotyping in crops like groundnut and rice. She has authored 20 research papers, 4 book chapters, and presented them at numerous national and international platforms. In recognition of her research excellence, she has been awarded the DST-SERB National Postdoctoral Fellowship and the Doreen Margaret Mashler Award..

R. BADDAM Rekha Baddam is a Senior Scientific Officer at ICRISAT, specializing in crop physiology. She holds an M.Sc. in Mathematics (2006) and a B.Sc. in Statistics (2002). With a strong analytical background and over a decade of experience, she contributes to research supporting agricultural sustainability and crop improvement. She is also a female drone pilot helping in research through drones.

V. MIKEŠ Ing. Vojtěch Mikeš is a doctoral candidate in Systems Engineering and Informatics at the Czech University of Life Sciences in Prague. His research is centered on the application of data science in the domain of plant phenotyping and stress detection. He is actively engaged in international collaborative projects, including partnerships with the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) in India and Çukurova University in Turkey, focusing on advanced analysis of 3D plant models and phenotypic traits under environmental stress. In 2024, he undertook a research stay at the University of Pisa, Italy, where he contributed to a project aimed at forecasting the occurrence of Sterility Mosaic Disease using high-throughput phenotyping data and computational modeling. Vojtěch possesses strong technical proficiency in tools and platforms such as Python, Spark, AWS, and Databricks. His interdisciplinary background combines expertise in systems engineering, informatics, and applied data analytics, positioning him at the intersection of computer science and agricultural research.

E. KÁNSKÁ Eva Kánská is an assistant professor at the Department of Information Technology Faculty of Economics and Management at Czech University of Life Sciences Prague. In her research activities, she focuses primarily on using information and communication technologies in agriculture, especially in data collection and analysis, information system design, and their application in practice. She is involved in more than seventeen research and educational projects, both at the national and international levels (Horizon 2020, Horizon Europe, Digital, Erasmus+ programs, etc.). Within these projects, she creates teaching materials, research activities, and professional practice cooperation. She is an active Czech Society for Information Technology in Agriculture (CSITA) member. An essential part of her pedagogical work is teaching in English, aimed at international students. She lectures on subjects related to information systems, programming, and digital technologies. In her teaching, she uses knowledge from current research and emphasizes connecting theory with practical use in the real environment.

6.3 Annotated 3D Point Cloud Dataset sada of Broad-Leaf Legumes Captured by High-Throughput Phenotyping Platform

GALBA, Alexander; MASNER, Jan; KHOLOVÁ, Jana; KARTAL, Serkan; STOČES, Michal et al. Annotated 3D Point Cloud Dataset of Broad-Leaf Legumes Captured by High-Throughput Phenotyping Platform. Online. *Scientific Data*. 2025, vol. 12, no. 1. ISSN 2052-4463. Dostupné z: https://doi.org/10.1038/s41597-025-06049-7. [cit. 2025-11-10].

www.nature.com/scientificdata

scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Annotated 3D Point Cloud Dataset of Broad-Leaf Legumes Captured by High-Throughput Phenotyping Platform

Alexander Galba[1], Jan Masner[1✉], Jana Kholová[1,2], Serkan Kartal[3], Michal Stočes[1], Vojtěch Mikeš[1], Pavel Šimek[1], Štěpánka Prokopová[1], René Fiala[1], Thorsten Karrer[4] & András Tóth[4]

This data descriptor presents novel, annotated 3D point cloud plant scans generated by a high-throughput phenotyping platform (LeasyScan, ICRISAT, India). It focuses on broad-leaf legume species (mungbean, common bean, cowpea, and lima bean). The dataset, generated by PlantEye(R) F600 technology, captures multispectral 3D scans of plant canopies. It includes 223 scans, providing detailed organ-level segmentation annotations for embryonic leaves, leaves, petioles, stems, and whole plants. The dataset fills a critical gap in plant phenomics research by offering a base of annotated data to support AI model development efforts in 3D computer vision. Data preprocessing, annotation procedures, and potential applications in crop research disciplines are further discussed. The dataset, preprocessing code, annotations, and a MIAPPE-compliant data sheet are also presented via the GitHub repository for further updates and expansion.

Background & Summary

Background. There is an increasing demand from plant-related research disciplines (e.g., crop breeding, gene banks, plant biologists, etc.), which require access to specific plant characters in large numbers of plants and with high precision and throughput. This has become possible with the development of sensor-based technologies (i.e., plant phenomics). These technologies typically generate vast amounts of data. However, the digital signal generated by the sensors requires data-processing algorithms to infer the desired plant features. Many of these algorithms are AI-based and require specific data inputs and pre-treatment (e.g., annotation) that are time- and resource-consuming to generate. To advance the development of plant traits inference algorithms in support of plant biology disciplines, there is a need to share the relevant datasets with the broad scientific community.

Related datasets. Not many public datasets provide annotated data in the form of 3D point clouds. A comprehensive list of public repositories can be found in the Papers with Code portal[1]. Another list is provided by Zifeng *et al.*[2]. Those datasets are mainly either LiDAR data generated by autonomous vehicles (complemented by 2D RGB image) or full indoor and outdoor scenes. There are only a few articles providing 3D point cloud plant scans (not available in standard repositories), such as soybean[3,4], rose[5], strawberry[6], or maize and tomato[7]. The mentioned datasets provide high-quality scans mostly generated by high-precision systems that do not allow high-throughput essays. To the best of our knowledge, no such an annotated dataset from a high throughput phenotyping platform as presented hereby is available publicly.

Provided plant species. In the presented dataset, we focus on broad-leaf legume species that have relatively simple canopy structures compared to other crop species. Namely, we provide the following species:

[1]Department of Information Technologies, Faculty of Economics and Management, Czech University of Life Sciences, Kamýcká 129, Prague, 165 00, Czech Republic. [2]International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru, 502 324, Telangana, India. [3]Department of Computer Engineering, Çukurova University, 01380, Adana, Türkiye. [4]Phenospex B. V., Jan Campertstraat 11, 6416 SG, Heerlen, The Netherlands. ✉e-mail: masner@pef.czu.cz

Name	Count
Total number of annotated scans	223
common bean	50
cowpea	45
lima bean	58
mungbean	71
Scans with all plants annotated using organs	141
Scans containing plants unannotated using organs	85
Scans containing some unannotated plants	3
Annotated classes	5
Annotated objects (all classes)	3 712
Annotated objects (Embryonic leaf)	1287
Annotated objects (Leaf)	1224
Annotated objects (Petiole)	814
Annotated objects (Stem)	88
Annotated objects (Plant)	299

Table 1. Summary of the dataset, including counts of annotated scans, species, plants, and classes. There are 223 scans (files) in total. Each scan contains 1–12 plants. Some plants could not be annotated using organ-level classes due to, e.g., wind distortion or overlapping. Instead, those plants were labeled by the Plant class (85 scans, 299 plants). 141 scans contain all plants annotated by organ-level classes.

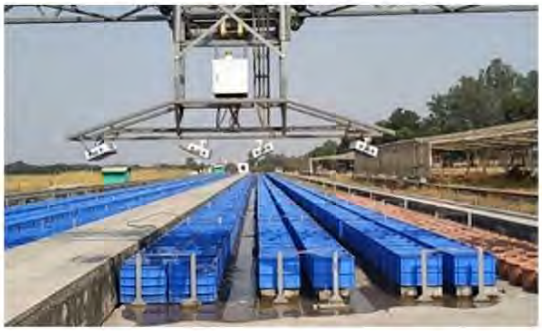


Fig. 1 The LeasyScan[9] high-throughput phenotyping platform used to gather the data. The picture shows the dual position (2 complementary, partially overlapping scanners capture the same area). The mounted scanners are moving over the field to capture the data (~2 500 m^2 area in 90 minutes).

mungbean (*Vigna radiata* L.), common bean (*Phaseolus vulgaris* L.), cowpea (*Vigna unguiculata* L.), and lima bean (*Phaseolus lunatus* L.). These have been generated as a part of crop improvement efforts at the International Crops Research Institute for Semi-arid Tropics (ICRISAT) and are dry-land grain legume crops – an important source of food and nutrition in semi-arid tropical agricultural systems.

Dataset summary. In summary, we provide annotated high-throughput plant scans in the form of 3D point clouds (*.PCD format). The counts of scans, species, objects, etc., are provided in Table 1. The dataset can be used for research not only in the field of plant phenomics but also generally in the development of 3D computer vision AI models that are currently far less developed than traditional 2D computer vision. The dataset is available at Figshare[8]. The provided dataset is annotated using the Segments.ai platform and can be easily re-imported into this software. All the code and data are also available as the GitHub (https://github.com/kit-pef-czu-cz/3d-point-cloud-dataset-plants) repository, which will be continuously updated with newly annotated data.

Methods

Technical equipment. The presented data were generated using a commercially available PlantEye technology (F600), which is a unique plant phenotyping sensor that combines a 3D scanner with multispectral imaging developed by Phenospex B. V. (PlantEye F600 multispectral 3D scanner for plants - PHENOSPEX). At the ICRISAT field (located in Hyderabad, India), during the data collection, the F600 scanners were mounted in a dual position (2 complementary, partially overlapping scanners capture the same area, illustrated in Fig. 1) and are set to cover the total cropped area of ~2 500 m^2 in 90 minutes. Details of the LeasyScan platform design can be found in[9].

The scanner captures plants' digital reflection in the form of two multi-spectral 3D point clouds where each point contains information on:

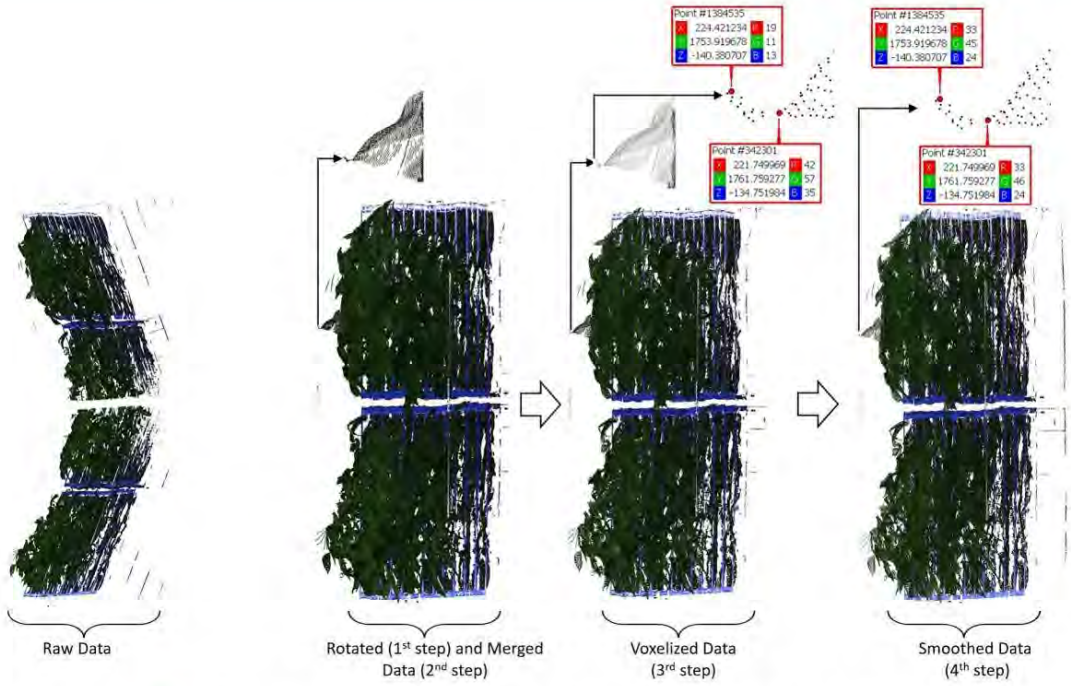


Fig. 2 Pre-processing steps of the raw scan files to extract only plant data for individual microplots.

- x, y, z coordinates in space
- Reflectance in Red, Green, Blue, and Near-Infrared spectra
- Reflectance of the 3D Laser (940 nm)

The 3D model of the plant is stored and pre-processed in proprietary software (HortControl) in an open *.PLY format. The files are accessible through a standard Breeding API (BrAPI) interface[10]. The 3D model of the plant can be used to directly measure or infer a range of plant parameters related to plant morphology and functions; at the moment, the inference algorithms are mostly limited to statistical-based prediction models.

Experiments. The hereby reported data comes from three experiments conducted in 2022 and 2023. Briefly, a single plant genotype was planted in one experimental unit ("microplot") consisting of a PVC tray (blue ones in Fig. 1) of 64 × 40 × 42.5 (length × width × height) cm containing ~50–60 kg of homogenized *Vertisols* collected from the ICRISAT farm. 12 seeds were planted in each tray and later thinned to 1–8 plants per tray, maintained throughout the vegetative growth phase. Plants were maintained up to ~ 35 days after planting, and the 3D point cloud data was obtained throughout the plant growth, typically twice a day. In each experiment, there were a multiple replications of each crop and genotype. At the LeasyScan platform, 24 microplots are grouped under a "barcode" area recognized by the scanning system and provided as a single raw scan. Each microplot has its identification based on a position within the barcode area (0_0 to 1_11).

Data preprocessing. The raw data obtained from the platform for each barcode is represented by the two scans (files) that are rotated to each other (illustrated on the left side of Fig. 2). The raw data are then pre-processed to extract only plant data for individual microplots. The initial step involves rotating the data to align flatly on the x-plane (see left side of Fig. 2). Both scans are merged into a single file in the second step. This merging process increases the point cloud density in the overlapping areas scanned by both scanners. Therefore, the third step involves a voxelization process[11] to rearrange the points in space uniformly. During the scanning process, certain point cloud data may be considered outlier values, where the color values differ significantly from the others. A smoothing process (4th step) was applied to unify these outliers in some point cloud data. In this process, each point data takes the average color value of the N nearest point data.

In the last step, we use a custom AI-based segmentation algorithm to separate plant data from background data such as soil and trays (details are the subject of another publication – please see the GitHub repository for details). The plant data are cut based on the fixed coordinates of each tray in the fifth step (Fig. 3). This step produces the input data for the annotation process.

The pre-processing was implemented independently to skip the coordinate-based cropping of soil data performed by the Phena pipeline in the HortControl software that operates the scanners. The current cropping is based on coordinates and, in some cases, cuts the plant data below the tray edge. We additionally implemented

Fig. 3 Last pre-processing step. On the left side, the zoomed-in plant data shows a single tray separated from the whole scan on the right side. Similarly, the data for all trays are separated from each other and saved as individual files for annotation.

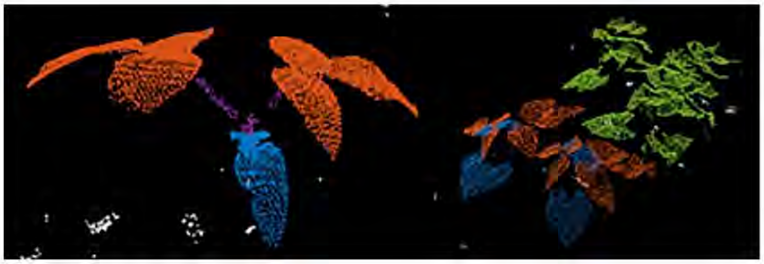


Fig. 4 Examples of annotated scans (orange color – Leaf, blue – Embryonic leaf, violet – Petiole, green - Plant). On the right side is a sample of a lower-quality scan on which it is impossible to recognize plant organs or individual plants.

the smoothing step, which is not part of the Phena pipeline. The pre-processing source code is provided to help work with the raw data.

Data annotation. The data were annotated using the online platform Segments.ai (https://segments.ai) under an academic license. Initial efforts included simultaneous drawing of cuboids (object detection) and segmentation for plant organs and whole plants. It was motivated by doing all possible annotations at once. However, this approach was too time-consuming. The Initial annotation of a single microplot took an average of two hours. It was also apparent that the segmentation drawing was less difficult than drawing cuboids. Annotation was, therefore, restricted to plant organs only. This reduced the time needed to annotate a file to an average of 30 minutes. *Plant* class segmentation annotations for all plants or cuboids for object detection (plants and organs) can be algorithmically generated using the segmentation annotations (see example code in the Usage Notes section).

There are 5 annotated classes within the dataset, specifically: *Embryonic leaf* (the juvenile leaves that are already present in the seed embryo and which have different morphology from other leaves), *Leaf*, *Petiole* (Leaf petiole), *Stem*, and *Plant*. The overlapping or distorted plants due to environmental conditions (e.g., wind) were additionally annotated using the *Plant* class (see right side of Fig. 4). Those unrecognizable plants naturally appear within the scans and cannot be avoided. In the provided dataset, each plant is either fully annotated by plant organs, annotated using the *Plant* class only, or unannotated. There are no partially annotated plants, only partially annotated scans that include unannotated plants (3 scans).

Data Records

The dataset has been deposited in Figshare[8]. All the shared data is structured into the following directories:

- Readme.md
  - Basic documentation for the dataset. Serves also as a description of the initial GitHub Repository.
- Data
  - Generated cuboid annotations
    - A folder that contains generated cuboids in .txt files using KITTI annotations format.
  - Point clouds
    - A folder containing all 3D point cloud files in .pcd format. The file naming convention is described in Table 2.
  - Segments-ai annotations.json
    - A file that contains segmentation annotations (organ-level mostly), where each point has an assigned class. The file is in the format from the Segments.ai platform (see *Segments-ai annotation format.md* for format description).
  - Segments-ai annotation format.md
    - A file that contains a description of the segments.ai annotation format.

Column name	Content description
Specie	Name of the plant specie that the file contains.
Exp. num.	Number of experiment, under which the scan was obtained at ICRISAT.
Bar code	Identification of a section within the experiment (position in the LeasyScan platform).
Tray	Identification of the tray within the section.
Date time	Timestamp of the scan in format *YYYYMMDDTHHMMSS*. The T is a divider.
Full-Part-Organs	"Full" determines, thar all organs of all plants in the scan were fully annotated. "Part", otherwise, means that the scan contains plant(s) where it is not possible to recognize their organs.
Full-Part-Plants	"Full" determines, whether all plants in the scan were annotated at least using the Plant class. Otherwise "Part". Part means, that in the scan, there are two or more plants that overlap so they cannot be distinguished from each other.
File name	Name of the file in the provided dataset. The name consist of the following columns, divided by dash ("-"): Exp. Num., Bar code, Tray, Date time.
Obj ID X	Multiple columns named Obj. ID X contains IDs of objects (annotated classes) that belongs to one plant.

Table 2. Description of the columns in annotation data.csv file that contains annotation records to assign annotated objects to individual plants.

	Average Precision		**F-Score @ best threshold**		**R^2**	**RMSE**
IoU	***0.3***	***0.5***	***0.3***	***0.5***	**count-only**	
Mean	0.701	0.317	0.759	0.478	0.788	2.153
Median	0.723	0.336	0.773	0.503	0.799	2.000
Best	0.796	0.389	0.817	0.554	0.905	1.491
Worst	0.546	0.162	0.659	0.323	0.601	2.944
Std. Dev.	0.071	0.071	0.046	0.068	0.085	0.450
Range	0.250	0.226	0.158	0.231	0.304	1.453
Var. Coeff.	10.1%	22.4%	6.1%	14.3%	10.8%	20.9%

Table 3. Summary of the results of the SECOND model for all outer and inner cross-validation combinations.

- Annotation data.csv (and .xlsx)
  - Annotations for plant organs to track their assignment to individual plants. A CSV file containing associations of annotated objects and individual plants in scan files. A single line in the file represents an individual plant and its organs. Table 2 provides a description of each column.
- Raw data.zip
  - Contains raw data from the scanner. There are always two files (each from a single scanner) for each bar code.
- MIAPPE_data.xlsx
  - Contains MIAPPE-compliant data sheet including mapping to the *Annotation data.csv* file.
- Code
  - Preprocessing from raw data
  - Cuboids generation
    - The folder contains an example code for generating cuboids for object detection for whole plants, together with the organs in the KITTI format. The folder contains an example annotation, an input point cloud file, and an output.
- Baseline evaluation
  - This folder contains full code and results for baseline evaluation using the SECOND and PointRCNN models with instructions on how to install, run, and reproduce the results.

Technical Validation

Plant scanning. The high-throughput phenotyping platform in ICRISAT was originally conceptualized in 2012 to detect key crop adaptations to environmental constraints (e.g., drought, heat, salinity) at the scale relevant to assist crop improvement programs[9]. In 2022, the platform was upgraded with the PlantEye F600 scanners, which has been used to generate data in the presented work. The LeasyScan fully automates the phenotyping process and creates insights into plant growth or changes in health for applications where detailed information or high numbers of plants are required, as referred in, e.g.,[12–14]. The PlantEye is built with high-quality standards to operate in any environment, such as growth chambers, labs, greenhouses, and fields. The technology is patented and widely used by scientists globally. For details about the scanner technology, refer to Phenospex website.

Annotation process. In order to validate the annotated data and minimize human errors, we created a protocol that included a double-checking process. Firstly, we trained every annotator and provided a detailed manual. Each file was assigned to a certain annotator. Another one was assigned to check the annotation first. The

	Average Precision		F-Score @ best threshold		R^2	RMSE
IoU	*0.3*	*0.5*	*0.3*	*0.5*	count-only	
Mean	0.544	0.258	0.709	0.480	0.718	3.615
Median	0.554	0.269	0.717	0.491	0.710	3.826
Best	0.692	0.366	0.809	0.588	0.847	2.384
Worst	0.385	0.158	0.595	0.374	0.580	4.835
Std. Dev.	0.076	0.059	0.054	0.060	0.065	0.703
Range	0.308	0.208	0.214	0.214	0.267	2.451
Var. Coeff.	14.0%	22.7%	7.6%	12.5%	9.0%	19.5%

Table 4. Summary of the results of the PointRCNN model for all outer and inner cross-validation combinations.

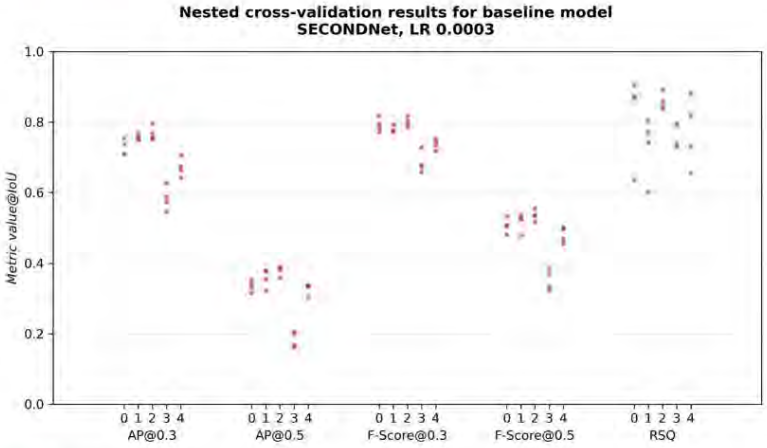


Fig. 5 Results visualization of the SECOND model (voxelization) evaluation for selected metrics. The X-axis shows different metrics for different outer test sets (0,1,2,3,4). Inner cross-validation combinations (train-validation) are represented by the dots. All results and their values can be found in the dataset[8] in the "Baseline Evaluation/Baseline Results" folder.

scan could either be returned to re-annotate or marked as checked. The checked scans were afterward checked again (marked as re-annotate or double-checked) by an expert or senior (experienced) annotator.

Baseline evaluation on object detection models. We conducted baseline experiments to assess the utility and applicability of the presented dataset using two standard object detection architectures: SECOND[15], which operates on voxel grids, and PointRCNN[16], which processes raw points. The codebase utilized the OpenPCDet library (https://github.com/open-mmlab/OpenPCDet) with minor modifications tailored to our dataset.

The dataset was randomly shuffled and partitioned into five-fold splits, each comprising 20%, enabling cross-validation. This resulted in training, validation, and test subsets in a 60:20:20 ratio. Models were trained using nested cross-validation (each fold is rotated as a test set; for each one, the remaining four are rotated as a validation set), ensuring a thorough evaluation and reducing biases related to fold selection, particularly beneficial given the relatively small size of the presented dataset[8].

Each model underwent training for up to 300 epochs, with an early stopping mechanism (patience of 100 epochs and warm-up of 25 epochs) based on the Average Precision (AP) metric at an Intersection-over-Union (IoU) threshold of 0.3. Default hyperparameters were used, with minor adjustments specific to dataset characteristics, including changes to the learning rate, detailed in the Baseline *evaluation/Code/OpenPCDet/tools/cfgs/README.md* file in the dataset repository[8].

Evaluation metrics included AP, F-Score, Precision, Recall, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the coefficient of determination (R^2, RSQ), providing a comprehensive view of model performance. Calculation methodologies for these metrics are explained in the supplementary Metrics notes. Results are summarized in Table 3 (SECOND) and Table 4 (PointRCNN), presenting descriptive statistics from each inner cross-validation combination evaluated across five different test sets. To illustrate the distribution across folds, selected metrics are visualized in Fig. 5 (SECOND) and Fig. 6 (PointRCNN). Comprehensive results for all metrics can be found in the dataset repository[8] under the *Baseline evaluation/Baseline results* folder.

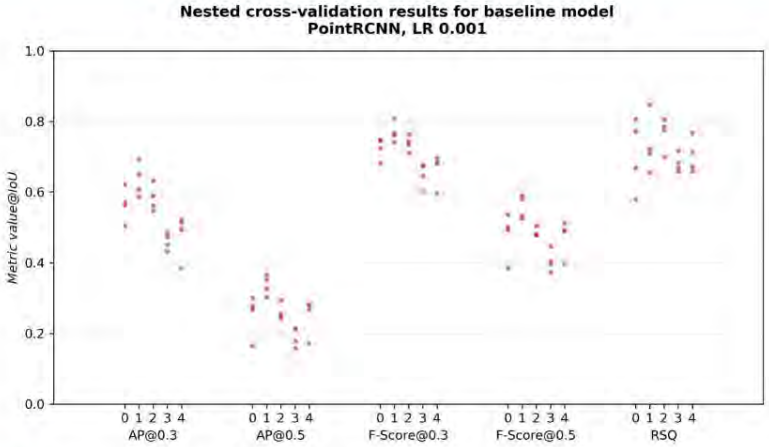


Fig. 6 Results visualization of the PointRCNN model (raw points) evaluation for selected metrics. The X-axis shows different metrics for different outer test sets (0,1,2,3,4). Inner cross-validation combinations (train-validation) are represented by the dots. All results and their values can be found in the dataset[8] in the "Baseline Evaluation/Baseline Results" folder.

Two main limitations were identified in regard to the data splitting and training procedures. First, Figs. 5, 6 highlight significant variations across different test sets. Employing a more sophisticated splitting strategy, considering plant numbers, scan size, or species, might yield more balanced results. Second, additional hyperparameter tuning could further enhance the models' performance.

Code availability

Together with the dataset[8], we first provide a sample code for preprocessing raw scan files to the format that is used for annotation. We also provide code for the automatic generation of cuboids for object detection tasks. Both codes take one file as input and output the result as another file. Third, we provide code for the model evaluations. The code is available in the Code directory.

Received: 31 January 2025; Accepted: 24 September 2025;
Published online: 10 November 2025

References

1. Machine Learning Datasets | Papers With Code. at https://paperswithcode.com/datasets?mod=point-cloud (2024).
2. Ding, Z. *et al.* Recent Advances and Perspectives in Deep Learning Techniques for 3D Point Cloud Data Processing. *Robotics 2023, Vol. 12, Page 100* **12**, 100 (2023).
3. Luo, L. *et al.* Eff-3DPSeg: 3D Organ-Level Plant Shoot Segmentation Using Annotation-Efficient Deep Learning. *Plant Phenomics* **5** (2023).
4. Sun, Y. *et al.* Soybean-MVS: Annotated Three-Dimensional Model Dataset of Whole Growth Period Soybeans for 3D Plant Organ Segmentation. *Agriculture 2023, Vol. 13, Page 1321* **13**, 1321 (2023).
5. Dutagaci, H., Rasti, P., Galopin, G. & Rousseau, D. ROSE-X: An annotated data set for evaluation of 3D plant organ segmentation methods. *Plant Methods* **16**, 1–14 (2020).
6. James, K. M. F., Heiwolt, K., Sargent, D. J. & Cielniak, G. Lincoln's Annotated Spatio-Temporal Strawberry Dataset (LAST-Straw). at https://arxiv.org/abs/2403.00566v1 (2024).
7. Schunck, D. *et al.* Pheno4D: A spatio-temporal dataset of maize and tomato plant point clouds for phenotyping and advanced plant analysis. *PLoS One* **16**, e0256340 (2021).
8. Galba, A. *et al.* Annotated 3D Point Cloud Dataset of Broad-Leaf Legumes Captured by High-Throughput Phenotyping Platform. at https://doi.org/10.6084/m9.figshare.28270742 (2025).
9. Vadez, V. *et al.* LeasyScan: A novel concept combining 3D imaging and lysimetry for high-throughput phenotyping of traits controlling plant water budget. *J Exp Bot*. https://doi.org/10.1093/jxb/erv251 (2015).
10. Selby, P. *et al.* BrAPI—an application programming interface for plant breeding applications. *Bioinformatics* **35**, 4147–4155 (2019).
11. Aleksandrov, M., Zlatanova, S. & Heslop, D. J. Voxelisation Algorithms and Data Structures: A Review. *Sensors (Basel)* **21**, 8241 (2021).
12. Sivasakthi, K. *et al.* Plant vigour QTLs co-map with an earlier reported QTL hotspot for drought tolerance while water saving QTLs map in other regions of the chickpea genome. *BMC Plant Biol* **18** (2018).
13. Tharanya, M. *et al.* Quantitative trait loci (QTLs) for water use and crop production traits co-locate with major QTL for tolerance to water deficit in a fine-mapping population of pearl millet (Pennisetum glaucum L. R.Br. *Theor Appl Genet* **131**, 1509–1529 (2018).
14. Sivasakthi, K. *et al.* Functional dissection of the chickpea (Cicer arietinum l.) stay-green phenotype associated with molecular variation at an ortholog of mendel's i gene for cotyledon color: Implications for crop production and carotenoid biofortification. *Int J Mol Sci* **20** (2019).
15. Yan, Y., Mao, Y. & Li, B. Second: Sparsely embedded convolutional detection. *Sensors (Switzerland)* **18** (2018).
16. Shi, S., Wang, X. & Li, H. PointRCNN: 3D object proposal generation and detection from point cloud. in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* **2019-June** (2019).

Acknowledgements
The results and knowledge included herein have been obtained owing to support from the following grants: Internal grant agency of the Faculty of Economics and Management, Czech University of Life Sciences Prague, grant no. 2023B0005 (Oborově zaměřené datové modely pro podporu iniciativy Open Science a principu FAIR); Ministry of Agriculture of the Czech Republic, grant number QK23020058 (Precision agriculture and digitization in the Czech Republic). We also want to thank the Segments.ai platform for the university license that was provided. Additional acknowledgments go to Anbazhagan Krithika, Sunita Choudhary, Baddam Rekha, and their students from ICRISAT for help with the initial data annotation protocol testing and the first round of annotations.

Author contributions
Alexander Galba – paper writing, code management, annotations generation; Jan Masner – conceptualization, paper writing, annotation management; Jana Kholová – conceptualization, data acquisition, paper revision; Serkan Kartal – data pre-processing; Michal Stočes – data management; Vojtěch Mikeš – data annotation; Pavel Šimek – paper revision, data management; Štěpánka Prokopová – data annotation; René Fiala – AI models development; Thorsten Karrer – data acquisition, pre-processing; András Tóth – data acquisition, pre-processing.

Competing interests
We hereby declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-06049-7.

Correspondence and requests for materials should be addressed to J.M.

6.4 Application of Quality Management System in the Research Process: A Case Study for Plant Phenotyping Research

GALBA, Alexander; KÁNSKÁ, Eva; MIKEŠ, Vojtěch; VANĚK, Jiří a JAROLÍMEK, Jan. Application of Quality Management System in the Research Process: A Case Study for Plant Phenotyping Research. Online. Agris on-line Papers in Economics and Informatics. 2024, roč. 16, č. 4, s. 79-86. ISSN 1804-1930. Dostupné z: https://doi.org/10.7160/aol.2024.160406. [cit. 2025-11-10].

Agris on-line Papers in Economics and Informatics

Volume XVI Number 4, 2024

Application of Quality Management System in the Research Process: A Case Study for Plant Phenotyping Research

Alexander Galba, Eva Kánská, Vojtěch Mikeš, Jiří Vaněk, Jan Jarolímek

Department of Information Technologies, Faculty of Economics and Management, Czech University of Life Science Prague, Czech Republic

Abstract

Phenomics research, driven by advancements in imaging and image processing, enables high-throughput measurements of plant traits, providing insights into growth, tissue development, and biochemical states. However, data accuracy is critical to reliable outcomes, especially in complex methods like 3D reconstruction and hyperspectral imaging. This study demonstrates the role of Quality Management Systems (QMS) in enhancing the research process in plant phenotyping. The study emphasizes the importance of a robust data quality assurance pipeline, focusing on error identification and improving data labeling processes through semi-automation. Root Cause Analysis (RCA) was employed to address discrepancies in annotated datasets and identify critical issues, such as misalignment in experimental protocols and operational errors, including the misplacement of irrigation hoses during data collection. Corrective actions, such as photo documentation and procedural revisions, significantly improved data quality. Additionally, algorithmic support streamlined the annotation process, increasing efficiency and data reliability. This integrated approach underscores the necessity of quality control in research, especially for geographically distributed teams working under variable conditions, and highlights the broader applicability of QMS in optimizing research outputs.

Keywords

Quality management system, data quality, plant phenotyping, research process, root cause analysis, data labeling process.

Galba, A., Kánská, E., Mikeš, V., Vaněk, J. and Jarolímek, J. (2024) "Application of Quality Management System in the Research Process: A Case Study for Plant Phenotyping Research", *AGRIS on-line Papers in Economics and Informatics*, Vol. 16, No. 4, pp. 79-86. ISSN 1804-1930. DOI 10.7160/aol.2024.160406.

Introduction

Phenomics research has benefited from the methods and devices allowing the high throughput measurements of plant traits at different levels, including growth-related traits, cells and tissue development, and biochemical and physiological states. Imaging and image processing developments enable capturing how these traits vary over time. Advancements in parallel and automated image acquisition allow for processing images of large plant populations under specific growth conditions. Image-processing capabilities enable 3D reconstruction of image data and automated quantification of biological features. (Simek et al., 2015) These advancements allow modeling at the systems level (Dhondt et al., 2013; Fiorani and Schurr, 2013; Kartal et al., 2021; Paulus, 2019; Pongpiyapaiboon et al., 2023; Sozzani et al., 2014; Vadez et al., 2015).

The goal of using digital methods is to achieve the best possible accuracy in identifying the biological parameters of the observed plants. A prerequisite for such accuracy is a high-quality database. Errors in data are the basis for poor-quality outputs or erroneous research conclusions. To achieve the best possible research results, it is necessary to pay attention to the quality of the entire research process (Ronanki et al., 2022).

The application of Quality Management Systems (QMS) in plant phenotyping research plays a pivotal role in ensuring data accuracy and consistency, especially with the increasing complexity of hyperspectral imaging and sensor-based techniques. For instance, a quality assurance pipeline developed for hyperspectral imaging systems ensures that spatial and spectral quality parameters are accurately maintained, enabling reliable detection of plant diseases such as Cercospora Leaf Spot through convolutional neural network (CNN)-supported data analysis. This quality-assured approach is crucial for evaluating and refining imaging systems in real-time plant health monitoring (Detring

et al., 2024). Another critical aspect of QMS in plant phenotyping is data management, essential for handling the vast amounts of digital data generated by automated sensing systems. Ensuring that data-sharing practices align with FAIR (Findability, Accessibility, Interoperability, and Reusability) principles can enhance the reproducibility and efficiency of research processes. (Ugochukwu and Phillips, 2022) The integration of advanced phenotyping methods, such as 3D reconstruction and leaf surface estimation, further highlights the need for robust data collection and analysis quality control to optimize workflows and reduce processing times (Daiki and Noshita, 2024).

The application of Quality Management Systems in the research process ensures consistent standards, minimizes errors, and enhances the reliability of outcomes. In clinical and public health research, QMS is essential to maintain protocol integrity, prevent deviations, and uphold the credibility of research findings. By implementing systematic procedures, research teams can ensure rigorous data quality and operational efficiency (Isere and Omorogbe, 2024). Additionally, in biomedical laboratories, QMS helps address quality issues, such as result stability and replication crises, by optimizing research processes and increasing effectiveness and efficiency. Despite initial resistance due to resource allocation and bureaucracy concerns, QMS offers significant advantages in process control and reliability.(Brünschwitz and Kleymann-Hilmes, 2024) Furthermore, operations research methods integrated with QMS in engineering and industrial contexts enable continuous improvement through mathematical modeling and optimization techniques, ensuring product and service quality across all phases (Parker, 2024). The application of QMS in research laboratories, such as in the petroleum industry, improves client satisfaction, reduces failure rates, and fosters industry-academia collaboration, providing a robust framework for process optimization and innovation (Vianna et al., 2022).

This article's main objective is to apply the quality management system in the research process to increase the quality of outputs.

Material and methods

Our research in the field of plant phenotyping is conducted by an international team located worldwide in various places and time zones. Researchers are located in the Czech Republic (The Czech University of Life Sciences), the Netherlands (Phenospex), India (ICRISAT - The International Crops Research Institute for the Semi-Arid Tropics), and Turkey (Cukurova University)

The research aims to estimate plant traits from 3D scans of plants acquired by the high-throughput phenotyping platform LeasyScan (built using Phenospex PlantEye F600 sensor). Subsequently, individual plant detection, organs, and other analyses are performed mainly using 3D computer vision methods. Obtaining 3D models outdoors brings efficiency to the entire research process. It does not require excessive manipulation with plants. Part of the process involves manually annotating datasets to train the artificial neural network models. The general scheme of the entire research is depicted in Figure 1.

The research commenced following the task assignment. Time sessions were set up for regular meetings where interim results were reviewed, schedules were discussed, and next steps were suggested. Control mechanisms were also part of the regular consultations. The method of selecting and checking a sample of data was chosen for control. The conclusions from the data control were regularly consulted, and measures were taken to correct the identified deficiencies.

The following procedure was chosen for the data flow: Data acquisition -> Data preprocessing -> Data Annotation. Data acquisition was performed in India using the LeasyScan platform. The source data was then stored and managed by Phenospex's Hortcontrol system, which is also available using BrAPI. Validated algorithms were used for

Source: Author

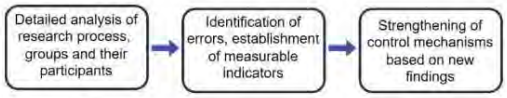
Figure 1: The general scheme of the entire research.

data preprocessing. The preprocessing consists of several steps, i.e., rotation, merging, voxelization, smoothing, and soil segmentation. The subsequent data annotation step was performed using the Segments.ai environment. All procedures were rechecked and verified (*Multi-Sensor Data Labeling Platform for Robotics and AV | Segments. Ai, n.d.; Phenospex - Smart Plant Analysis and Phenotyping Systems, n.d.*).

Over time, it became clear that the set rules needed to be revised to ensure the overall quality of the research process. Weaknesses in the quality of the data, rather than in the quality of the steps used, were identified. Regular online meetings of all groups and information sharing, including the control mechanisms, are still needed to eliminate the deficiencies.

The primary source of deficiencies was identified in the data annotation process for generating test and control datasets using artificial neural networks. The number of annotated plants for the ground-truth did not match the predictions. The first step to correcting this situation was to increase the qualification of persons designated for data labeling through regular training and more detailed checks. It turned out that this procedure only solved the consequences and not the source of the error. It was evident that it was necessary to revise the entire research process and find the source of the inaccuracies.

The control mechanisms, consisting of the selection of samples and their control, needed to be revised, and it was decided that it was necessary to find a way to check all the acquired data. At the same time, it was essential to check all steps to exclude the possibility of multiple sources of errors. To avoid further problems in the future, the methodological procedure expressed in Figure 2 Plan to improve the research process was chosen.



Source: Author

Figure 2: Plan to improve research process.

It was important for the whole process to choose a method to identify the bottleneck. During the discussion, it became clear that the methods used are mainly focused on the commercial area or processes in the public sector. Their application in research is only partially obvious. Most methods focus on streamlining the process regarding time, costs, or risk analysis. Some methods are extensive, and their application would require longer (ISO). A common attribute of these methods is creating a detailed procedure diagram in various formats. This procedure is then supplemented with tools and procedures to detect deficiencies and thereby increase the quality of the entire process. The diagram expresses the sequence of causes and effects and the influence of the main actors. (Bali et al., 2021; *ISO 9001:2015 - Quality Management Systems - Requirements, n.d.*; Serrat, 2017; Shook, 2008; Starzyńska and Hamrol, 2013; Venkatasubramanian et al., 2003) The display allows a better understanding of the sequence of events and the identification of critical points. The methods then focus on these and increase quality by applying appropriate control mechanisms. It is a suitable tool to express the quality of the process in a measurable quantity. In practice, this may be time, the number of erroneous outputs, or a combination of similar values. Based on such a criterion, the success of the applied method can then be evaluated. An overview of the methods is shown in Figure 3 (Tarí and Sabater, 2004a).

The assessment of which method is appropriate depends on the perspective of the participants and the definition of the target parameters. Finally, the point of view of the greatest possible generality in use and applicability for scientific research was chosen. From this point of view, the root case analysis method was chosen. A Root Cause Analysis (RCA) is the process of identifying the most fundamental reason for the problem, which, if eliminated or corrected, would prevent the problem from reoccurring. Various techniques for analysis can be used for this method, and therefore, its application is suitable for multiple environments, including high-risk industries such as medicine or drug development (Andersen and Fagerhaug, 2000; Percarpio et al., 2008; Andersen, B. and Fagerhaug, T. (2006); Wilson et al., 1993; Yuniarto, 2012).

Creating a measurable parameter expressing the source data's quality and accuracy proved essential. In addition to the position in 3D space, the data from the device also contained information about the point's color. A method based on the height parameter of individual points (z-axis value) was chosen for the check. The points in each data sample were divided into 2 groups according to height. The division had the following assumptions:

The seven basic quality control tools	The seven management tools	Other tools	Techniques
Cause and effect diagram	Affinity diagram	Brainstorming	Benchmarking
Check sheet	Arrow diagram	Control plan	Departmental purpose
Control chart	Matrix diagram	Flow chart	analysis
Graphs	Matrix data analysis method	Force field analysis	Design of experiments
Histogram	Process decision	Questionnaire	Failure mode and effects analysis
Pareto diagram	Programme chart	Sampling	Fault tree analysis
Scatter diagram	Relations diagram		Poka yoke
	Systematic diagram		Problem solving methodology
			Quality costing
			Quality function deployment
			Quality improvement teams
			Statistical process control

Source: Tarí and Sabater, 2004b

Figure 3: Overview of the methods.

- Group 1 may contain points belonging to the color of soil and flower pots
- Group 2 may contain points in a color belonging only to plants

Subsequently, the color in individual groups was checked algorithmically for all samples. It turned out that in Group 2, there were points whose color did not correspond to the points of plants. This finding was supplemented by the knowledge from the manual labeling of samples, where some points and shapes could not be labeled as parts of plants. The number of points not corresponding to plants in group 2 was determined as a measurable data quality criterion. This finding focused the research team's attention on preparing and implementing experiments when obtaining source data. In addition to creating a quantifiable indicator expressing the quality of the data, we focused on the entire research process.

Results and discussion

To apply RCA in our research, we created a process diagram. Our problem was the quality of the source data. After the analysis, we identified 4 areas that impact the source data. These were the technologies used, the settings of individual experiments, the execution of separate experiments, and the data processing methods. In these areas, any shortcomings could affect the quality of the data. We set up control mechanisms to verify the correct functioning in all areas. Some control mechanisms are determined by the type of area. For technical equipment, this meant performing a setting check and calibration. To set up individual experiments, we conducted discussions with everyone involved in the settings. In the area of experiment execution, we conducted more detailed operator training. For processing methods, we prepared control outputs of the algorithms used to verify their functioning. However, we were unable to improve the quality of the data. After a detailed analysis of the experiment execution, it was found that the operators needed to adhere sufficiently to the established procedures. To rectify this situation, a requirement for photographic time documentation of each experiment was introduced. The documentation showed that the irrigation hoses (black color) were carelessly placed, causing their points to overlap with the source data (Figure 4).

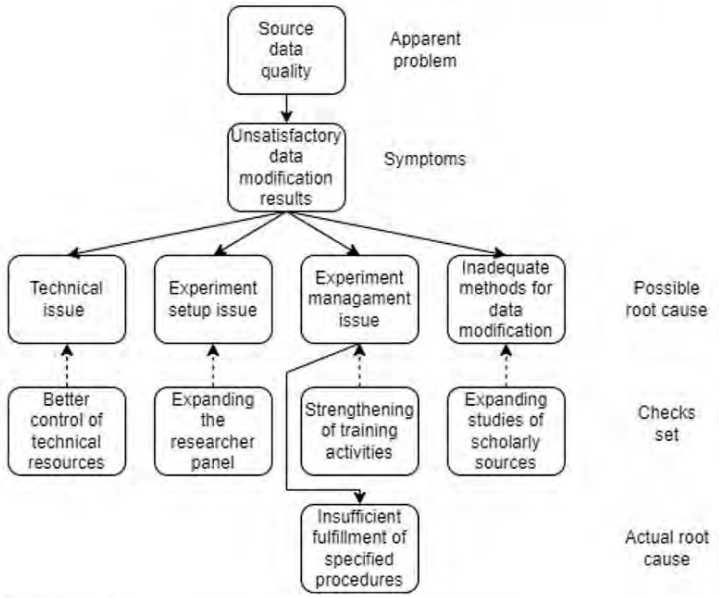
Source: Author

Figure 4: Irrigation hoses.

After correcting errors in the execution of experiments and implementing photo documentation control, the source data quality increased. After corrective actions, we selected and reviewed a new data set of acquired data. It no longer showed anomalies in the 2nd group of points. The entire RCA tree diagram is shown in Figure 5.

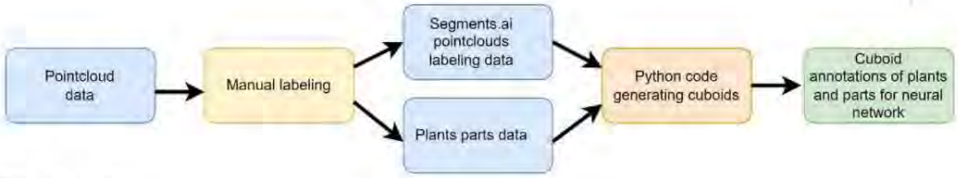
Taking photo documentation and storing it for each experiment brought success by revealing the cause of the insufficient data quality. After discussing why the entire process of conducting experiments needed to be set up from the beginning, including photo documentation, it turned out that the reason was technical complexity. It was necessary to install a photo device for the outdoor scanner and connect it to the data storage. This measure seemed expensive and unnecessary.

Finding the error also resulted in another discovery: unsatisfactory data labeling speed and labeling data quality. We discovered that trained operators still need help accurately labeling the designated class of objects. After an analysis, we changed the marking procedure so that the operators' duty was to mark only defined categories of plant organs (leaves, bay leaves, stems, and petiole) and to record the belonging of the parts to a specific plant in the form of recording identifiers in numerical terms. An algorithm was subsequently written to mark the whole plant, automatically generating this mark. Further streamlining of the labeling process was found in dropping the labeling of object detection points and replacing this process with an algorithm that produces this labeling automatically from labeled parts. These changes made it possible to include a larger group of operators in the labeling process, thereby speeding up the entire labeling process. The set procedure is shown in Figure 6.



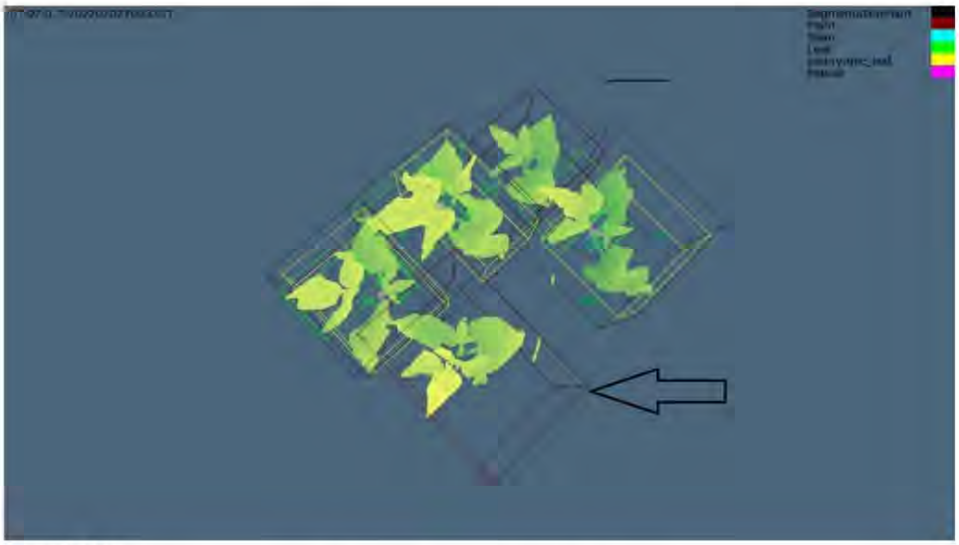
Source: Author

Figure 5: Root cause analysis tree diagram.



Source: Author

Figure 6: Semi-automated labeling process.



Source: Author

Figure 7: Visual control of labeling data.

Automatic cuboid generation also proved helpful in detecting incorrect labeling. The labeling process was supplemented with images of the generated cuboids, enabling more efficient control. Visual identification of suspicious samples was much more accessible. See Figure 7.

Conclusion

The mentioned research project is part of broader research in plant phenotyping using artificial intelligence. Our findings showed that using critical insight in the research process is very beneficial. The quality of research outputs directly depends on the data quality used. Our research takes place in an international team, geographically separated and in different time zones. The nature of the experiments performed does not allow their exact repetition. Cultivation of plants for experimental purposes in an outdoor environment cannot be repeated with identical conditions. All this makes it challenging to identify the sources of errors and inaccuracies. Creating a measurable parameter to express data quality was a guide to finding the source of errors. In the outdoor plant scanning environment, the acquired data is less rich than in the case of indoor scanning. Therefore, the highest possible quality of the source data was achieved important. The use of optimization methods for improving quality is mainly concentrated in the area of the process in the sphere of production, trade, and services. Their use in a research environment is more challenging. The hierarchy of research teams is not strictly defined, and the results often depend on individual abilities, ideas, and creativity. The fact that we were not satisfied with the original quality of the data and implemented procedures to increase the quality of the entire process moved us to a higher level. At the same time, we increased the efficiency of part of the process (labeling), and this shortened the time frame and expanded the database for generating higher-quality research outputs. An important factor was also the fact that we introduced measures in source data quality in the following areas.

Acknowledgements

The results and knowledge included herein have been obtained owing to support from the following institutional grant. Internal grant agency of the Faculty of Economics and Management, Czech University of Life Sciences Prague, grant no. IGA 2023A0017.

This work was supported by the EC's Horizon Europe funding in the project CODECS grant agreement No. 101060179.

This work was supported by the EC's Digital Europe Programme in the project AGRITECH EU grant agreement No. 101123258.

Corresponding author:
Doc. Ing. Pavel Šimek, Ph.D.
Department of Information Technologies, Faculty of Economics and Management
Czech University of Life Sciences Prague,
Kamýcká 129, 165 00 Prague - Suchdol, Czech Republic
E-mail: simek@pef.czu.cz

References

[1] Andersen, B. and Fagerhaug, T. (2000) "*Root cause analysis: simplified tools and techniques*", pp. 155. [Online]. Available: https://books.google.com/books/about/Root_Cause_Analysis.html?hl=csandid=i_EJAQAAMAAJ. [Accessed: Sept. 8, 2024]. ISBN 0873894669.

[2] Andersen, B. and Fagerhaug, T. (2006) "*Root Cause Analysis, Second Edition: Simplified Tools and Techniques*" , 240 p., Quality Oress. ISBN 9780873896924.

[3] Bali, P., Kutikuppala, L. V. S., Avti, P. and Medhi, B. (2021) "*Data Fraud and Essence of Data Verifiability, Quality Assurance Implementation in Research Labs*", pp. 137-159. ISBN 978-981-16-3074-3. DOI 10.1007/978-981-16-3074-3_9.

[4] Brünschwitz, S. and Kleymann-Hilmes, J. (2024) "Benefits and approaches of a quality management system in biomedical research laboratories", *Bundesgesundheitsblatt, Gesundheitsforschung, Gesundheitsschutz*, Vol. 67, No. 1, pp. 99-106. E.ISSN ISSN 1437-1588. DOI 10.1007/S00103-023-03797-Y.

[5] Daiki, S. and Noshita, K. (2024) "A comparative study of plant phenotyping workflows based on three-dimensional reconstruction from multi-view images", *BioRxiv*. DOI 10.1101/2024.03.21.586185.

[6] Detring, J., Barreto, A., Mahlein, A.-K. and Paulus, S. (2024) "*Quality Assurance of Hyperspectral Imaging Systems for Neural Network supported Plant Phenotyping*". DOI 10.21203/RS.3.RS-4648326/V1.

[7] Dhondt, S., Wuyts, N. and Inzé, D. (2013) "Cell to whole-plant phenotyping: the best is yet to come", *Trends in Plant Science*, Vol. 18, No. 8, pp. 428-439. ISSN 1360-1385. DOI 10.1016/J.TPLANTS.2013.04.008.

[8] Fiorani, F. and Schurr, U. (2013) "Future Scenarios for Plant Phenotyping", *Annual Review of Plant Biology*, Vol. 64, pp. 267-291. ISSN 1543-5008. DOI 10.1146/annurev-arplant-050312-120137.

[9] Isere, E. E. and Omorogbe, N. E. (2024) "Quality Management in Clinical and Public Health Research: A Panacea for Minimising and Eliminating Protocol Deviations in Research Operations", *Nigerian Medical Journal*, Vol. 65, No. 3, pp. 367-375. ISSN 0300-1652. DOI 10.60787/NMJ-V65I3-421.

[10] ISO 9001:2015 - Quality management systems - Requirements. (n.d.) [Online]. Available: https://webstore.ansi.org/standards/iso/iso90012015 [Nov. 15, 2023].

[11] Kartal, S., Choudhary, S., Masner, J., Kholova, J., Stoces, M., Gattu, P., Schwartz, S. and Kissel, E. (2021) "Machine Learning-Based Plant Detection Algorithms to Automate Counting Tasks Using 3D Canopy Scans", *Sensors,* Vol. 21, No. 23, pp. 8022. ISSN 1424-8220. DOI 10.3390/S21238022.

[12] Multi-sensor data labeling platform for robotics and AV | Segments.ai. (n.d.). [Online]. Available: https://segments.ai/. [April 28, 2024].

[13] Parker, J. (2024) "The application of operation research in the quality management engineering", *Advances in Operation Research and Production Management*, Vol. 1, No. 1, pp. 25-32. ISSN 1687-9155. DOI 10.54254/3006-1210/DIRECT/2532.

[14] Paulus, S. (2019) "Measuring crops in 3D: Using geometry for plant phenotyping", *Plant Methods*, Vol. 15, No. 1, pp. 1-13. ISSN 1746-4811. DOI 10.1186/S13007-019-0490-0/FIGURES/5.

[15] Percarpio, K. B., Watts, B. V. and Weeks, W. B. (2008) "The effectiveness of root cause analysis: what does the literature tell us?", *Joint Commission Journal on Quality and Patient Safety*, Vol. 34, No. 7, pp. 391-398. ISSN 1553-7250 . DOI 10.1016/S1553-7250(08)34049-5.

[16] Phenospex - Smart plant analysis and Phenotyping systems. (n.d.). [Online]. Available: https://phenospex.com/ [Nov. 15, 2023].

[17] Pongpiyapaiboon, S., Tanaka, H., Hashiguchi, M., Hashiguchi, T., Hayashi, A., Tanabata, T., Isobe, S. and Akashi, R. (2023) "Development of a digital phenotyping system using 3D model reconstruction for zoysiagrass", *The Plant Phenome Journal*, Vol. 6, No. 1, ISSN 2578-2703. DOI 10.1002/PPJ2.20076.

[18] Ronanki, S., Pavlík, J., Masner, J., Jarolímek, J., Stočes, M., Subhash, D., Talwar, H. S., Tonapi, V. A., Srikanth, M., Baddam, R. and Kholová, J. (2022). An APSIM-powered framework for post-rainy sorghum-system design in India, *Field Crops Research*, Vol. 277, p. 108422. ISSN 0378-4290. DOI 10.1016/J.FCR.2021.108422.

[19] Serrat, O. (2017) "The Five Whys Technique", In: *Knowledge Solutions*, pp. 307-310. Springer, Singapore. ISBN 978-981-10-0982-2. DOI 10.1007/978-981-10-0983-9_32.

[20] Shook, J. (2008) "Managing to Learn: Using the A3 Management Process to Solve Problems, *Gain Agreement, Mentor and Lead"*, pp. 10-11. ISBN 1934109207.

[21] Šimek, P., Stočes, M., Vaněk, J., and Masner, J. (2015) "Mobile accessibility of information sources in the areas of agriculture, forestry, water management, food industry and rural development", *Agrarian Perspectives XXIV: Global Agribusiness and the Rural Economy*, pp. 440-446. ISBN 978-80-213-2581-4.

[22] Sozzani, R., Busch, W., Spalding, E. P. and Benfey, P. N. (2014) "Advanced imaging techniques for the study of plant growth and development", *Trends in Plant Science*, Vol. 19, No. 5, pp. 304-310. ISSN 1878-4372. DOI 10.1016/J.TPLANTS.2013.12.003.

[23] Starzyńska, B. and Hamrol, A. (2013) "Excellence toolbox: Decision support system for quality tools and techniques selection and application", *Total Quality Management and Business Excellence*, Vol. 24, No. 5-6, pp. 577-595. ISSN 0954-4127. DOI 10.1080/14783363.2012.669557.

[24] Tarí, J. J. and Sabater, V. (2004a) "Quality tools and techniques: Are they necessary for quality management?", *International Journal of Production Economics*, Vol. 92, No. 3, pp. 267-280. ISSN 0925-5273. DOI 10.1016/J.IJPE.2003.10.018.

[25] Ugochukwu, A. I. and Phillips, P. W. B. (2022) "Data sharing in plant phenotyping research: Perceptions, practices, enablers, barriers and implications for science policy on data management", *Plant Phenome Journal*, Vol. 5, No. 1. ISSN 2578-2703. DOI 10.1002/PPJ2.20056.

[26] Vadez, V., Kholová, J., Hummel, G., Zhokhavets, U., Gupta, S. K. and Hash, C. T. (2015) "LeasyScan: a novel concept combining 3D imaging and lysimetry for high-throughput phenotyping of traits controlling plant water budget", *Journal of Experimental Botany*, Vol. 66, No.18, pp. 5581. ISSN 0022-0957. DOI 10.1093/JXB/ERV251.

[27] Venkatasubramanian, V., Rengaswamy, R., Yin, K. and Kavuri, S. N. (2003) "A review of process fault detection and diagnosis: Part I: Quantitative model-based methods", *Computers and Chemical Engineering*, Vol. 27, No. 3, pp. 293-311. ISSN 0098-1354. DOI 10.1016/S0098-1354(02)00160-6.

[28] Vianna, E. L. F., De Figueiredo, V. V., Da Silva, C. M. F., Bertolino, L. C. and Spinelli, L. (2022) "Impact of implementing quality control systems in laboratories associated with teaching and research institutions – The case study of the laboratory for macromolecules and colloids in the petroleum industry. International Journal of Metrology and Quality Engineering, Vol. 13, pp. 4. ISSN 2107-6847. DOI 10.1051/IJMQE/2022004.

[29] Wilson, P. F. ., Dell, L. D. and Anderson, G. F. (1993) "*Root cause analysis: a tool for total quality management*", pp. 216. ASQC Quality Press, 1993. ISBN 9780873891639.

[30] Yuniarto, H. A. (2012) "The Shortcomings of Existing Root Cause Analysis Tools", *Proceedings of the World Congress on Engineering 2012 Vol III,* WCE 2012, July 4 - 6, 2012, London, U.K. ISSN 2078-0958.

6.5 Forecasting Sterility Mosaic Disease in Pigeonpea Using Dynamic Bayesian Networks and 3D Point Cloud High-throughput Scanning Platform

MIKEŠ, Vojtěch; KOCIAN, Alexander; KHOLOVÁ, Jana; MASNER, Jan; KLECZKOWSKI, Adam; SHARMA, Mamta; CHESSA, Stefano**; GALBA, Alexander;** ŠIMEK, Pavel. Forecasting Sterility Mosaic Disease in Pigeonpea Using Dynamic Bayesian Networks and 3D Point Cloud High-throughput Scanning Platform. Online. In: *2025 21st International Conference on Intelligent Environments (IE)*. IEEE, 2025, s. 1-8. Dostupné z: https://doi.org/10.1109/ie64880.2025.11130066. [cit. 2025-11-10].

Forecasting Sterility Mosaic Disease in Pigeonpea Using Dynamic Bayesian Networks and 3D Point Cloud High-throughput Scanning Platform

Vojtěch Mikeš*†, Alexander Kocian†, Jana Kholová‡*, Jan Masner*
Adam Kleczkowski§, Mamta Sharma‡, Stefano Chessa†, Alexander Galba* Pavel Šimek*

*Department of Information Technologies, Czech University of Life Sciences Prague, Prague, Czech Republic
†Department of Computer Science, University of Pisa, Pisa, Italy
‡International Crops Research Institute for the Semi-Arid Tropics: Patancheru, Hyderabad, Telangana, India
§Mathematics and Statistics, University of Strathclyde, Glasgow, Great Britain

***Abstract*—This paper explores how high-throughput phenotyping can be integrated with machine learning models to efficiently forecast the Sterility Mosaic Disease using a small amount of training data. This approach is generalized through the use of a Dynamic Bayesian Network (DBN). To predict the spread of the virus, the entire network is decomposed into several distributed and cooperative learning modules. The EM algorithm is used to learn the parameters for each module. Upon iterative convergence, the estimated hidden state vector of one module serves as input control for the next. The parameter estimates of the final module are used to formulate a predictor capable of forecasting q-days ahead.**

To demonstrate the effectiveness of the proposed DBN, its performance is evaluated using real-world data from ICRISAT, Patancheru, Hyderabad, Telangana, India. Physiological data was collected using 3D point cloud technology, while environmental data was recorded by a local weather station.

***Index Terms*—Pathology, Machine learning, Expectation Maximization, Cooperation, Distributed Learning**

I. Introduction

The global demand for food is projected to rise by over 50%, potentially even more when factoring in the effects of global warming and recent pandemic outbreaks [1]. This growing need places increased pressure on cultivating high-yielding, nutrient-rich crops, such as Pigeonpea (Cajanus cajan), one of the most widely cultivated crops worldwide. Currently, Pigeonpea crops cover approximately 5.4 million hectares, primarily in India, with an annual global yield of around 4.5 million tons [2].

Pigeonpea is highly valued for its nutritional benefits, serving as a vital food source for humans and livestock in tropical and subtropical regions where other crops struggle to thrive due to low water availability [2].

The results and knowledge included herein have been obtained owing to support from the following grant: Internal grant agency of the Faculty of Economics and Management, Czech University of Life Sciences Prague, grant no. 2023A0017. Moreover, this work is funded partially within the project AGRITECH Spoke 9 - Codice progetto MUR: AGRITECH "National Research Centre for Agricultural Technologies" - CUP CN00000022, of the National Recovery and Resilience Plan (PNRR) financed by the European Union "Next Generation EU", and the project funded by Scottish Funding Council's International Science Partnerships Fund.

However, Pigeonpea production faces a significant challenge in the form of Sterility Mosaic Virus (SMV). This virus is one of the most critical threats to crop production, causing Mosaic Sterility Disease (SMD) with symptoms such as yellow spots on the upper leaves, stunted growth, and plant sterility (among other effects) [3]. SMV has resulted in global yield losses exceeding 300 million USD [4].

Predicting the evolution of SMD is essential for safeguarding Pigeonpea production, ensuring food security, and promoting sustainable agricultural practices. It empowers stakeholders with the tools needed to mitigate the virus' impact and adapt to future challenges.

Prediction models integrated into phenotyping systems, such as those used in plant breeding, could significantly enhance the classification of plant resistance to SMV. These models have the potential to accelerate the breeding process and reduce associated costs by minimizing the reliance on manual measurements, which are often error-prone [5], [6].

The evolution of disease assessment in agriculture has seen significant advancement over the years. Initially, manual disease assessment relied heavily on visual inspections by experts, which were time consuming, labor intensive, and prone to human error [7]. With the advent of 2D scanning methods, high-resolution images of plant leaves could be captured and analyzed using machine learning algorithms. These methods improved accuracy and efficiency, allowing for early detection of diseases and reducing the reliance on chemical treatments.

The next leap in technology came with the development of 3D point-cloud high-throughput scanners. These advanced systems use techniques such as LiDAR and multispectral imaging to create detailed 3D models of plants [8]. By capturing the intricate structure of plants, these scanners provide comprehensive data that can be analyzed to monitor plant

health and detect diseases with unprecedented precision [9]. This evolution not only enhances disease detection, but also supports sustainable agriculture by enabling precise interventions and optimizing resource use. In addition, this approach can help reduce the costs associated with plant phenotyping, plant breeding, and phytopathology.

Related studies have explored various methods for predicting and managing SMD in Pigeonpea. The study in [10] analyzed the incidence of SMV across four locations in India using field data collected during six growing seasons. The authors demonstrated that hybrid models, which combine Support Vector Regression (SVR) and ARIMA, generally outperformed standalone methods [10] [11], and proved to be efficient in regions where the seasonal mean severity of SMD exceeded 1%. In contrast, standalone models like SVR performed better for lower incidence values. Additionally, their findings emphasized the spatial variability in the correlation between the incidence of SMV and weather parameters such as temperature, rainfall, and humidity.

In another study [12], pre-trained Convolutional Neural Network (CNN) architectures were used to detect SMV by visual classification of healthy and infected leaves. The training dataset consisted of images from field experiments, including infected and healthy Pigeonpea leaves. This approach achieved an average accuracy of 88%, highlighting the potential of AI-driven methods for early disease detection. However, all of the aforementioned models require large amounts of training data, which can be resource intensive and time consuming [13].

Predicting plant disease over time is difficult due to the complex, dynamic, and multifaceted nature of plant health. The challenges include insufficient and noisy data, environmental variability, complex interactions between multiple factors, and the difficulty of obtaining longitudinal data. Despite these challenges, advances in machine learning, sensor technologies, and data collection methods are gradually improving the ability to predict plant diseases more accurately.

Rather than considering the chain from weather data to virus infection as one machine learning problem, we model it as Dynamic Bayesian Netowrk (DBN), which can be decomposed into several distributed and cooperative learning modules, whose parameters are learned from physiological data sensed by the 3D point clout scanning platform and weather-related control data by using the Expectation Maximization (EM) algorithm. At iterative convergence, the trained parameters are used to formulize a predictor that is capable of forecasting the SMD several days in advance.

II. Methodology

A. The 3D PointCloud

The LeasyScan platform [14] provides data for plant-related variables. It uses the Phenospex PlantEye F600 multispectral 3D scanner (Figure 1), which is capable of determining and exporting plant characteristics non-destructively [15]. All data sets are stored in a breeding database and accessed through BrAPI, making them readily available for additional analysis. BrAPI is a project that aims to maintain interoperability among various breeding databases.

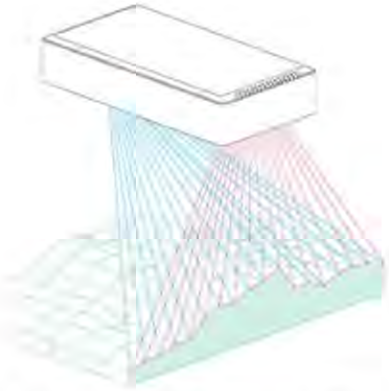


Fig. 1. PlantEye F600 used for high-throughtput plant scanning. The image shows how the reconstruction of the canopy (green) is done with use of scanner laser technology (red). [15]

Some of the measurements used here are calculated by the LeasyScan platform from 3DPC data such as digital biomass or leaves penetration depth. However, some of the measurements do not use 3DPC data for their calculation, for example, the height of the plant, which is calculated as $plant\ height = D_p - pot\ height$ where D_p is the distance from the scanner to the top of the plant [14].

These plant traits, but not all, are used for the creation of the predictor model. LeasyScan platform is considered a high-throughput scanning platform, which means that it is capable of producing a large amount of data in a short time (around 4800 experimental sectors that can be identified in less than two hours with the help of barcodes) with respect to [14].

All plant-dependent (state data) measurements listed below have been created by the LeasyScan platform. The scanner is able to create 3-D point cloud scans of plants and calculate morphological traits that are used for further data processing. The list is not exhaustive of all possible variables from LeasyScan as some variables, e.g. helper buckets for the ndvi, npci and psri indexes, are dropped and only the average of these indexes is considered. All volume variables are in mm^3 and all areas are in mm^2:

- Digital biomass
- Average greenness
- Plant height
- Plant MAX height
- Leaf angle
- Leaf area projected
- Leaf area
- Leaf area index
- Leaf inclination
- Light penetration depth
- Average NDVI
- Average NPCI
- Average PSRI

However, for the sake of the SMD forecasting, not all sensed plant attributes are helpful. Having uncorrelated features in

the feature vector ensures that a Dynamic Bayesian Network learns efficiently, generalizes well, and avoids redundant computations. The result is improved optimization, reduced overfitting, and aligns with the principles of effective feature selection [16]. After data cleaning, the final list of variables is shown below. We will see that these measures are sufficient for short-term SMD forecasting.

- Digital biomass
- Plant height
- Light penetration depth

The selected experiment that was used as the basis for the prediction of the SMD was carried out in Pigeonpea (Cajanus cajan (L.) Millsp.) crops at ICRISAT, Patancheru, Hyderabad, Telangana, India, from 30 July 2023 to 15 September 2023. In total, four genotypes were used in the experiment:

- BDN-1
- ICP 2376
- ICP 7035
- ICP 8863

Genotypes were separated into control and inoculated groups. Only Pigeonpea crops from the inoculated group were selected for the experiment.

The crops were planted in pots with a maximum of two plants per pot. The pots were designed in a grid system that was labeled with a unique barcode. The plants were inoculated with SMV on 14 August 2023 and ended on 5 September 2023. After that, the plants were monitored twice a day by the ICRISAT disease team, who made manual measurements, which were then processed and scanned twice a day by the LeasyScan scanner.

The data set contains 6612 raw measurements. These measurements are then processed in a way that only the daily average of measurement is present in the data set. This processing step reduced the data to 3600 measurement values. After data cleansing like dropping null values or deleting corrupted data, the data set ended up with 1680 respective measurement values. The descriptive statistics of the final data set are shown in Table I, showing the size, minimum, maximum, and mean value, standard deviation, as well as the quartiles Q1,...Q3. The barcodes were then assigned to the data according to the LeasyScan specifications described in [14]. Each barcode describes an experiment comprising 4 plants that are measured separately in a time span of 18 days. Every two or three days, a measurement point has been added to the data frame. The missing data between has been interpolated.

For the environmental parameters, the data were obtained from meteorological stations three kilometers from the actual location of the experimental fields in ICRISAT. The descriptive statistics are shown in Table II Some environmental data points were missing. As the weather in Hyderabad during this time of the season is generally stable with a periodic day-night pattern, we decided to impute any missing values using data from the public Weather Underground database [17].

TABLE I
PHYSIOLOGICAL PARAMETERS OF THE PLANTS RECORDED WITH LEASYSCAN PLATFORM USED FOR THE SMD FORECAST SIMULATIONS

Measure	Digital Biomass (mm^3)	Height (mm)	Depth of light pen. (mm)
Mean (μ)	9.5×10^6	238.6	161.7
Std. Dev. (σ)	8.2×10^6	74.1	52.3
Min	7.4×10^3	65.9	0.4
Max	3.9×10^7	475.1	369.9
Q1	3.6×10^6	185.6	130.7
Q2 (Median)	6.6×10^6	233.3	155.9
Q3	1.3×10^7	286.1	190.8

TABLE II
ENVIRONMENTAL DATA RECORDED DURING THE TIME OF THE EXPERIMENT

Measure	Temperature (°C)	Humidity (%)
Mean (μ)	27.070	72.573
Std. Dev. (σ)	1.763	9.356
Min.	22.600	57.100
Max.	30.400	93.800
Q1 (1st quartile)	25.850	65.150
Q2 (Median)	27.350	71.900
Q3 (3rd quartile)	28.225	79.625

Finally, the ground truth data set is composed of a manually measured disease index (%). A common approach is to use the weighted average of vigor, greenness and severity of the disease [18] described in Table III. For each barcode, ground-truth measurements were performed once every two or three days during the time of the experiment.

TABLE III
STERILITY MOSAIC VIRUS SEVERITY RATING SCALE FOR THE PIGEONPEA EXPERIMENT

Variable	Scale	Description
Vigur	0 - 100	100 is scored when inoculated plants have equivalent vigour and healthiness compared to non-inoculated control plants of same genotype
Greenness	1 - 5	Value of 5 means plant is fully green and 1 means plant is yellowish
Disease severity	0 - 100	Value 100 means completely infected and 0 means no sign of infection

B. System model and architecture

Botanical epidemiology strongly depends on statistical models. The relationship between the epidemic component and environmental parameters is non-linear. A simple approach that yet captures the underlying processes, is to use a generalization of the monomolecular model describing the evolution of infection as a function of time [19]

$$I(t) \propto \kappa \left(1 - e^{-\mu f(t-t_0)}\right) \quad (1)$$

where κ is the upper limit of the response, $f(t)$ describes the response of the pathogen for a given temperature, and μ is

the rate of growth, which can be related to, for example, the wetness duration, t_0 is a parameter describing the time-offset. Note that $f(t)$ is explicitly a function of time.

For a simple linear case, $f(t - t_0) = t - t_0$ in the exponent of (1), and the evolution in (1) is the solution to a first-order linear ordinary differential equation.

$$\frac{dI(t)}{dt} + \mu I(t) = \mu\kappa. \tag{2}$$

Sampling with rate $1/\Delta$ and rearranging the result, we obtain

$$I_t = (1 - \mu\Delta)I_{t-1} + \mu\Delta\kappa. \tag{3}$$

The difference equation in (3) represents the progression of the leaf infection as a dynamic linear model expressed in the form of a time series. Let I_t be modeled as a Markov process of the first order that cannot be observed directly but only through noisy observations. For the sake of simplicity, we normalize the sampling rate in (3). Then, a state-space system that represents the difference equation of the form [20]:

$$\begin{aligned} \boldsymbol{h}_t &= (\boldsymbol{I} - \boldsymbol{A}\ \boldsymbol{G}_t\})\,\boldsymbol{h}_{t-1} + \boldsymbol{B}g_t + \boldsymbol{n}_t; \boldsymbol{n}_t \propto \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma_n}) \\ \boldsymbol{I}_t &= \boldsymbol{C}\boldsymbol{h}_t + \boldsymbol{w}_t; & \boldsymbol{w}_t \propto \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma_w}) \\ \boldsymbol{h}_1 &= \boldsymbol{\mu}_1 + \boldsymbol{n}_1; & \boldsymbol{n}_1 \propto \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_1) \end{aligned} \tag{4}$$

with the symbol $\boldsymbol{G}_t \triangleq \mathrm{diag}\{\boldsymbol{g}\}_t, t = 1, \ldots, T$. Here, $\boldsymbol{A}$ is the state matrix; $\boldsymbol{B}$ is the input matrix; $\boldsymbol{C}$ is the output matrix. Additive noise $\mathcal{N}(-; \boldsymbol{m}, \boldsymbol{\Sigma})$ is assumed to be Gaussian with mean $\boldsymbol{m}$ and covariance $\boldsymbol{\Sigma}$. Moreover, the K-dim column vector $\{\boldsymbol{g}_t : \boldsymbol{g}_t \in \mathbb{R}^K, t \in [1, T]\}$ acts as (deterministic) control vector related to the environmental data. The (noisy) measurement vector $\{\boldsymbol{I}_t : \boldsymbol{I}_t \in \mathbb{R}^D, t \in [1, T]\}$ contains sensor data with D dimensions. The noisy and hidden state $\{\boldsymbol{h}_t : \boldsymbol{h}_t \in \mathbb{R}^K, t \in [1, T]\}$ captures the patterns or context of a sequence in a summary vector, which can provide insight into the evolution of infection. The joint state-measurement distribution has the form

$$p\left(\boldsymbol{h}_{1:T}, \boldsymbol{I}_{1:T}\right) = p\left(\boldsymbol{h}_1\right) p\left(\boldsymbol{I}_1 | \boldsymbol{h}_1\right) \prod_{t=2}^{T} p\left(\boldsymbol{h}_t | \boldsymbol{h}_{t-1}\right) p\left(\boldsymbol{I}_t | \boldsymbol{h}_t\right). \tag{5}$$

The DBN in the lower half of Fig. 2 is based on the factorization of the conditional distributions in (5). In the diagram, each edge represents a conditional dependency, and each node represents one of the three types of variables.

Let us focus on the control data once more. Once the leaves are infected, Equ. (1) suggests that temperature and humidity are the sole driving forces for the progression of infection. However, we will see that the process is more complicated. Specifically, further infection strongly depends on the current physiological state of the plant. Following this approach, let the control vector $\boldsymbol{g}_t$ in (4) be composed of information related to digital biomass $b(\theta(t), p(t))$, the height of the plant $l(\theta(t), p(t))$, and the penetration depth of light $\Phi(\theta(t), p(t))$, where $\theta(t)$ and $p(t)$ represent the temperature of the air and the humidity of the air as a function of time, respectively. These

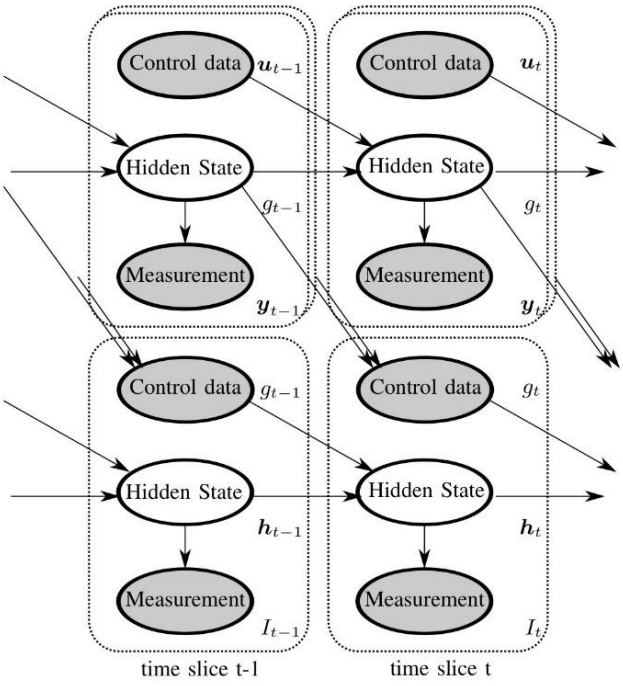


Fig. 2. 2-time-slice "Deep" Dynamic Bayesian Network where the hidden state of one network acts as control of an other.

physiological data have their own dynamics, often according to an exponential evolution [21], [22].

Thus, we model the physiological data as another Markov process of first order that cannot be observed directly but only through noisy observations by the 3D point cloud $\boldsymbol{y}_t$, $t = 1, \ldots, T$ and controlled (deterministically) by $\boldsymbol{u}_t$ comprising the time series for air temperature and air humidity. Following this approach, the state-space system for each physiological variable has the form [23], [24], [25]:

$$\begin{aligned} \boldsymbol{g}_t &= \boldsymbol{A}'\boldsymbol{g}_{t-1} + \boldsymbol{B}'\boldsymbol{u}_{t-1} + \boldsymbol{\zeta}_t; & \boldsymbol{\zeta}_t \propto \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}'_{\boldsymbol{n}}) \\ \boldsymbol{y}_t &= \boldsymbol{C}'\boldsymbol{g}_t + \boldsymbol{\eta}_t; & \boldsymbol{\eta}_t \propto \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}'_{\boldsymbol{w}}) \\ \boldsymbol{g}_1 &= \boldsymbol{\mu}'_1 + \boldsymbol{\zeta}_1; & \boldsymbol{\zeta}_1 \propto \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}'_1) \end{aligned} \tag{6}$$

Note that $\boldsymbol{y}_t$ is non-periodic with missing data for most time points. Each DBN in the upper half of Fig. 2 corresponds to the factored state measurement distribution based on (6). We have decomposed a complex learning problem into a simple distributed and cooperative learning problem. Additional state-space systems that model more complex evolutions, such as Elman networks with sigmoid or tanh non-linearity [26] can be easily added.

Figure 3 provides a high-level view of the proposed system architecture. Physiological data and weather data are sensed by LeasyScan as well as by the local weather station at ICRISAT, respectively. Data has been pre-processed, transformed into a suitable format and stored in a MySQL database. Each

DBN whose state-space is matched to the system response of any of the physiological variables (exponential, exponential saturation, sigmoid, tanh, Elman Neural Network, etc.) tracks the respective data and forwards the result to the second layer of the deep DBN. The latter is responsible for tracking the final disease variable. The network parameters of the DBN can then be used to produce an accurate prediction of the disease several days in advance.

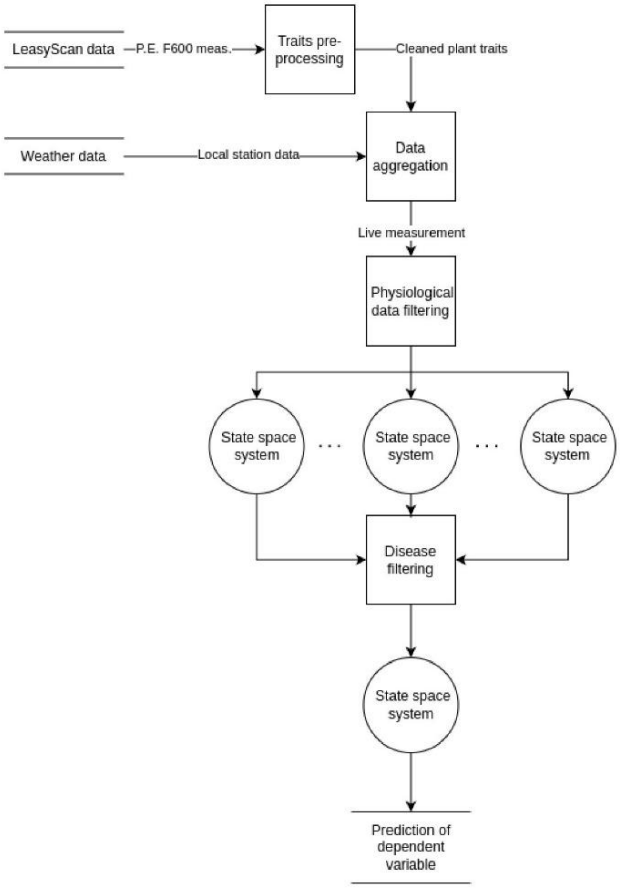


Fig. 3. High-level view of the proposed system architecture

C. Model estimation and prediction

We use an iterative method to estimate the parameter vector for either state-space system. We start with that in (1). This involves hypothesizing the unobserved (missing) data, denoted as $\boldsymbol{h}_{1:T}$. Starting from iteration $i = 0$, the E-step of the algorithm calculates the expected log-likelihood $p\left(\boldsymbol{h}_{1:T}, \boldsymbol{y}_{1:T}|\boldsymbol{\theta}\right)$ based on the observed data $\boldsymbol{y}_{1:T}$ and the current parameter estimate $\boldsymbol{\theta}^{[i]}$:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{[i]}) = \mathbb{E}\{\ln p\left(\boldsymbol{h}_{1:T}, \boldsymbol{y}_{1:T}|\boldsymbol{\theta}\right)|\boldsymbol{y}_{1:T}, \boldsymbol{\theta}^{[i]}\}. \tag{7}$$

For a dynamic Bayesian network (DBN), using the joint probability from the model, this becomes:

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{[i]}) =& \mathbb{E}\left\{\ln p\left(\boldsymbol{h}_1|\boldsymbol{\theta}\right)|\boldsymbol{y}_{1:T}, \boldsymbol{\theta}^{[i]}\right\} \\ &+ \sum_{t=2}^{T} \mathbb{E}\left\{\ln p\left(\boldsymbol{h}_t|\boldsymbol{h}_{t-1}, \boldsymbol{\theta}\right)|\boldsymbol{y}_{1:T}, \boldsymbol{\theta}^{[i]}\right\} \\ &+ \sum_{t=1}^{T} \mathbb{E}\left\{\ln p\left(\boldsymbol{y}_t|\boldsymbol{h}_t, \boldsymbol{\theta}\right)|\boldsymbol{y}_{1:T}, \boldsymbol{\theta}^{[i]}\right\}. \end{aligned} \tag{8}$$

In the M-step, the parameter $\boldsymbol{\theta}^{[i+1]}$ is updated to maximize the expected log-likelihood:

$$\boldsymbol{\theta}^{[i+1]} = \arg\max_{\theta} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{[i]}). \tag{9}$$

The sequence of log-likelihood values $\left\{\ln p\left(\boldsymbol{y}_{1:T}|\boldsymbol{\theta}^{[i]}\right)\right\}_{i=0}^{\infty}$ converges steadily to a stationary point of $\ln p\left(\boldsymbol{y}_{1:T}|\boldsymbol{\theta}\right)$. [27], [28]. For more details on the derivation of the EM-based tracking algorithm, refer to [24], [20].

The EM algorithm alternates between prediction and correction. When the feedback loop is interrupted, the algorithm can still make predictions on its own without receiving any new responses. Using this idea, we create a predictor that forecasts q-steps ahead. Starting from the state evolution described in Equation (1), we calculate the expected value based on the most recent observation and the steady-state parameter vector, which comes from the latest estimate provided by the EM algorithm. Hence, mean and covariance of the free running q-step predictor for the last stage in (1), we obtain after a few straightforward algebraic operations, it follows for the q-time step free-running predictor of the measurement data that

$$I_{T+q}^{[\infty]} = \boldsymbol{C}^{[\infty]}\left(\boldsymbol{I} - \boldsymbol{A}^{[\infty]}\boldsymbol{G}_T\right)\boldsymbol{h}_{T+q-1}^{[\infty]} + \boldsymbol{C}^{[\infty]}\boldsymbol{B}^{[\infty]}\boldsymbol{g}_T. \tag{10}$$

with error covariance

$$\boldsymbol{V}_{T+q}^{[\infty]} = \boldsymbol{C}^{[\infty]}\boldsymbol{\Xi}_q\left(\boldsymbol{C}^{[\infty]}\right)^T \tag{11}$$

with the short-cut

$$\boldsymbol{\Xi}_q \triangleq \left(\left(\boldsymbol{I} - \boldsymbol{A}^{[\infty]}\boldsymbol{G}_T\right)\boldsymbol{V}_{T-1+q}\left(\boldsymbol{I} - \boldsymbol{A}^{[\infty]}\boldsymbol{G}_T\right)^T + \boldsymbol{\Sigma}_{\boldsymbol{n}}^{[\infty]}\right) \tag{12}$$

The other state-space systems are constructed analogously.

D. Initialization Issues

The EM algorithm is sensitive to its initialization [28], as different starting points can lead to different stationary points of the log-likelihood function. To guide the EM algorithm towards a possible global maximum of the log-likelihood function, we rely on the structure of the time series for the disease in (1).

We start with the state-space system in (4) that has a system response according to exponential saturation. Suppose that there exist three training points $y_{t_1}^{(\mathrm{obs})}, y_{t_2}^{(\mathrm{obs})}, y_{t_3}^{(\mathrm{obs})}$ at time

TABLE IV
OVERVIEW OF SELECTED BARCODES AND ITS SMD INCIDENCES FOR SIMULATIONS

Barcode	Plant #	max SMD %	min SMD %
204-11-2	4	90	60
204-4-1	1	10	5

instants t_1, t_2, t_3, respectively, located within the first training samples $T_{\min}$. Inserting the observation points into (1), it follows for the system parameters that

$$\begin{aligned} \mu &= \frac{\log(1 - y_{t_1}^{(\mathrm{obs})}/\kappa) - \log(1 - y_{t_3}^{(\mathrm{obs})}/\kappa)}{t_3 - t_1}, \\ t_0 &= \frac{1}{\mu}\left(\log(1 - y_{t_1}^{(\mathrm{obs})}/\kappa)\right) + t_1, \\ \kappa &= \frac{y_{t_2}^{(\mathrm{obs})}}{1 - \exp\{-\mu(t_2 - t_0)\}} \end{aligned} \tag{13}$$

If the logarithms in (13) exist, we can iterate through the set of equations until convergence is achieved. It follows from (1) and (3), for the state matrix $\boldsymbol{A}^{[0]}$ that

$$\boldsymbol{A}^{[0]} = \mu^{\star}\ \mathrm{diag}\{\tfrac{1}{T}\textstyle\sum_{t=1}^{T} \boldsymbol{u}_t\}^{-1}.$$

and hence for the control matrix $\boldsymbol{B}^{[0]}$ that

$$\boldsymbol{B}^{[0]} = \boldsymbol{A}^{[0]}\kappa^{\star}.$$

Otherwise, the system matrices are filled with positive random values $\varepsilon \ll 1$. The measurement matrix $\boldsymbol{C}^{[0]}$ is initialized as the all-one matrix divided by the column width of the matrix. Moreover, the noise covariance matrices are initialized as $\boldsymbol{\Sigma}_{\boldsymbol{n}}^{[0]} = \boldsymbol{\Sigma}_{\boldsymbol{w}}^{[0]} = \boldsymbol{\Sigma}_1^{[0]} = \varepsilon \boldsymbol{I} \ll 1$. Finally, the initial state

$$\boldsymbol{\mu}_1^{[0]} = \boldsymbol{C}^{\dagger} y_{t_1}^{(\mathrm{obs})} \tag{14}$$

where $\boldsymbol{C}^{\dagger}(\cdot)$ denotes the Moore-Penrose inverse of the argument.

The other state-space system in (6) has a system response according to an exponential function. For the sake of simplicity, we initialize all system matrices with random values as described above.

III. EXPERIMENTAL RESULTS AND DISCUSSION

To demonstrate the effectiveness of the proposed DBN, we evaluate its performance using two different barcodes. One barcode corresponds to a plant that regularly develops SMD over time, while the other represents a plant that only remains slightly infected. The experiment lasted $T_{\max} = 18$ days according to the procedure outlined in Section II-A. Table IV provides a summary of the data set.

In the first experiment, we evaluated the SMD in (%) versus days of inoculation. Fig. 4 shows the performance of the DBN. For barcodes 204-11-2 and 204-4-1, the measured SMD values are marked as a circle and a triangle, respectively. The DBN was trained during the first five days after inoculation of the plant according to Section II-D, following the parametric approach. As time progresses, the EM algorithm learns the parameters from the sensor data. The tracked data at iterative convergence is plotted as a dashed line. The forecast value of the 1-step free-running predictor in (10) - (12) is recorded as solid line. For barcode 204-11-2 that resembles exponential

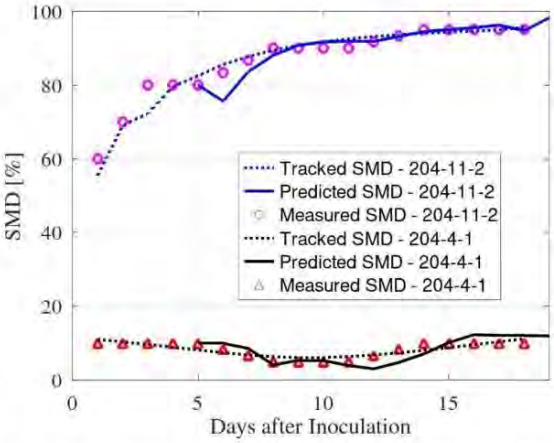


Fig. 4. Forecast of SMD. Each barcode represents a different resilient plant. Forecast is done with $q = 1$

saturation, it can be seen that with increasing time, the predictor achieves accurate performance over the entire range of SMD levels, though only four initial values are available, For barcode 204-4-1 that exhibits valley-type behavior, the evolution of training data is flat. Hence, the DBN randomly initializes the system parameters according to Section II-D. It can be seen that our predictor is still capable of providing decent quality data.

So far, we have considered a prediction length of $q = 1$. It would be interesting to see how our predictor performs as q increases. This parameter dictates how many data points the model omits before generating subsequent predictions. During this experiment, we distributed the training data in the first time samples $T_{\min}$, subject to the first predicted disease $I_{T_{\min}+q}^{[\infty]}$ within the valid range of $]0, 100]\%$. Table V lists the minimum pilot length $T_{\min}$ for different barcodes and prediction length q. It can be seen that for Barcode 204-

TABLE V
SELECTED q VALUES FOR BARCODES USED IN MAE CALCULATION

q	T_{min} for 204-11-2	T_{min} for 204-4-1
1	5	5
2	5	7
3	6	7
4	5	7
5	8	9
6	7	8
7	6	8

11-2 with exponential behavior, between 5 and 8 training points are sufficient for the EM algorithm to achieve iterative

convergence. Barcode 204-4-1 with valley behavior, up to 9 samples are required. though. This behavior is mainly caused by the fact that the trend becomes clear, once training points are located at the bottom of the valley of the disease curve. Fig. 5 shows the mean absolute error (MAE) in per cent points according to

$$\mathrm{MAE}(q) = \frac{1}{N} \sum_{T=T_{\min}}^{T_{\max}} |I_{T+q} - I_{T+q}^{[\infty]}| \tag{15}$$

as a function of the prediction length q for different barcodes over all $N = T_{\max} - T_{\min} + 1$ samples. It can be seen that

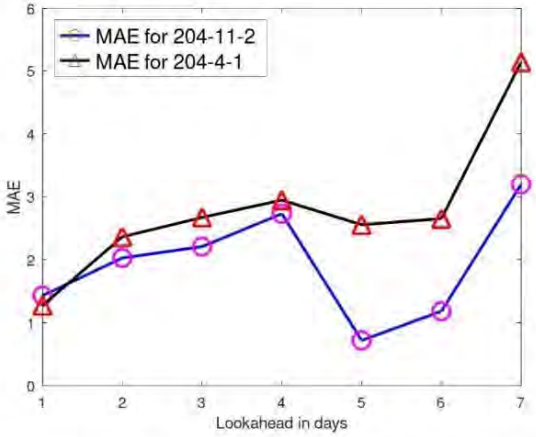


Fig. 5. Mean Absolute Error in (15) for the Barcodes 204-11-2 and 204-4-1 as a function of the prediction length.

the MAE is limited by 6 % points throughout the time sample range N.

Fig. 6 reports the measured versus predicted SMD for the Pigeonpea with the prediction length q as parameter. The 1:1 line has been added as dashed line for the sake of convenience. To see how well our model predicts an outcome, we added the coefficient of determination R^2

$$R^2(q) = 1 - \frac{\sum_T \left(I_{T+q} - I_{T+q}^{[\infty]}\right)^2}{\sum_T \left(I_{T+q} - \overline{I_{T+q}}\right)} \tag{16}$$

to the plot. Note that the symbol $\overline{(\cdot)}$ denotes the mean value of the argument. For $q = 1$, the trend line has a slope of 1.182 and an intercept point of 0.294 ($R^2 = 0.94$). For $q = 3$, the behavior is very similar, that is $R^2 = 0.99$. Analogous can be said for other prediction lengths, which have been omitted from the already overcrowded plot.

We found that the DBN is capable of predicting SMD even if the evolution of the disease does not match the model. However, without proper initial conditions, the model may be incapable of fitting the training data, rendering future forecasts impossible, or resulting in excessive divergence. When the evolution of initial training points matches those

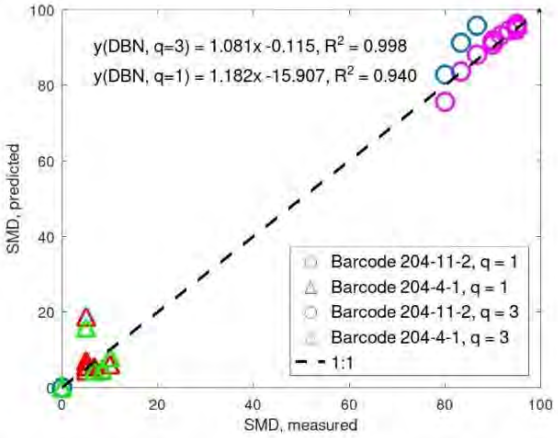


Fig. 6. Relation between measured and predicted SMD values for PigeonPea with the prediction length as parameter. The dashed line corresponds to the 1:1 relationship.

of the underlying DBN model, the parametric approach in (13) leads to valid initial system parameters. When no data structure can be exploited, the initial system parameters can only be randomly initialized at the expense of accuracy.

IV. Conclusions

This study highlights the successful integration of high-throughput phenotyping with machine learning models for forecasting the spread of Sterility Mosaic Virus using limited training data. The proposed approach, based on a Dynamic Bayesian Network (DBN), demonstrates the potential to model complex biological processes efficiently and with reduced labor requirements.

The experiment carried out at ICRISAT showcased the effectiveness of combining physiological data from a three-dimensional point cloud scanning platform with environmental data from a local weather station. The decomposition of the DBN into distributed and cooperative learning modules allowed for modular optimization tailored to specific system responses, such as exponential growth, saturation, and sigmoid behavior.

The use of the Expectation-Maximization (EM) algorithm ensured accurate parameter estimation for each module, with iterative convergence enabling seamless integration between modules. This design facilitated the development of a robust q-step (day) ahead predictor, capable of providing reliable forecasts.

The evaluation of the DBN on two experiments involving four plants over 18 days confirmed its effectiveness in capturing the dynamics of SMD spread. These findings underscore the versatility and efficiency of the DBN framework, paving the way for broader applications in agricultural disease forecasting and environmental monitoring.

While DBNs are versatile and effective for modeling sequential and temporal data, their limitations in scalability, assumptions, and computational requirements must be carefully addressed. Strategies such as approximate inference, hybrid models, and advanced learning techniques can mitigate some of these challenges, but they often come with trade-offs in complexity and accuracy.

REFERENCES

[1] A. Hassoun, S. Jagtap, H. Trollman, G. Garcia-Garcia, N. A. Abdullah, G. Goksen, F. Bader, F. Ozogul, F. J. Barba, J. Cropotova, P. E. Munekata, and J. M. Lorenzo, "Food processing 4.0: Current and future developments spurred by the fourth industrial revolution," Food Control, vol. 145, p. 109507, 3 2023.

[2] E. O. Fatokimi and V. A. Tanimonure, "Analysis of the current situation and future outlooks for pigeon pea (Cajanus Cajan) production in Oyo State, Nigeria: A Markov Chain model approach," Journal of Agriculture and Food Research, vol. 6, p. 100218, 12 2021.

[3] B. R. Sayiprathap, A. K. Patibanda, V. Prasanna Kumari, K. Jayalalitha, H. K. Ramappa, E. Rajeswari, L. Karthiba, K. Saratbabu, M. Sharma, and H. K. Sudini, "Salient Findings on Host Range, Resistance Screening, and Molecular Studies on Sterility Mosaic Disease of Pigeonpea Induced by Pigeonpea sterility mosaic viruses (PPSMV-I and PPSMV-II)," Frontiers in Microbiology, vol. 13, p. 838047, 4 2022.

[4] L. Manjunatha, H. K. Ramappa, A. Puyam, and N. Srinivasa, "Pigeonpea Sterility Mosaic Virus: a threatening virus of pigeonpea, current scenario and its control," Indian Phytopathology, vol. 74, pp. 885–891, 12 2021.

[5] C. H. Bock, J. G. A. Barbedo, E. M. Del Ponte, D. Bohnenkamp, and A.-K. Mahlein, "From visual estimates to fully automated sensor-based measurements of plant disease severity: status and challenges for improving accuracy," Phytopathology Research 2020 2:1, vol. 2, pp. 1–30, 4 2020.

[6] K. S. Chiang and C. H. Bock, "Understanding the ramifications of quantitative ordinal scales on accuracy of estimates of disease severity and data analysis in plant pathology," Tropical Plant Pathology, vol. 47, pp. 58–73, 2 2022.

[7] P. Yadav and P. Singh, "Disease detection techniques in plants: Transition from manual to automation," in New Approaches for Multidimensional Signal Processing (R. Kountchev, R. Mironov, and K. Nakamatsu, eds.), (Singapore), pp. 93–109, Springer Nature Singapore, 2023.

[8] Q.-Y. Zhou, J. Park, and V. Koltun, "Open3D: A modern library for 3D data processing," arXiv:1801.09847, 2018.

[9] F. Okura, "3d modeling and reconstruction of plants and trees: A cross-cutting review across computer graphics, vision, and plant phenotyping," Breeding Science, vol. 72, no. 1, pp. 31–47, 2022.

[10] R. K. Paul, S. Vennila, S. K. Yadav, M. N. Bhat, M. Kumar, P. Chandra, A. K. Paul, and M. Prabhakar, "Weather based forecasting of sterility mosaic disease in pigeonpea (Cajanu cajan) using machine learning techniques and hybrid models," The Indian Journal of Agricultural Sciences, vol. 90, pp. 1952–1958, 12 2020.

[11] R. Kumar Paul, S. Vennila, N. Singh, P. Chandra, S. Yadav, O. Sharma, S. Nisar, M. Bhat, M. Rao, and M. Prabhakar, "Seasonal Dynamics of Sterility Mosaic of Pigeonpea and its Prediction using Statistical Models for Banaskantha Region of Gujarat, India," Journal of the IndIan SocIety of agrIcultural StatIStIcS, vol. 72, no. 3, pp. 213–223, 2018.

[12] S. Yashwant Pawar and D. Hingole, "Exploring artificial intelligence technique for detection of pigeon pea sterility mosaic disease," The Pharma Innovation Journal, no. 12, pp. 482–489, 2023.

[13] B. R. Sayiprathap, A. K. Patibanda, M. Mantesh, S. Hiremath, N. Sagar, C. N. L. Reddy, C. R. Jahir Basha, S. E. Diwakar Reddy, M. Kasi Rao, R. M. Nair, and H. K. Sudini, "Sterility mosaic disease of pigeonpea (cajanus cajan (l.) huth): Current status, disease management strategies, and future prospects," Plants, vol. 13, no. 15, 2024.

[14] V. Vadez, J. Kholová, G. Hummel, U. Zhokhavets, S. Gupta, and C. T. Hash, "LeasyScan: a novel concept combining 3D imaging and lysimetry for high-throughput phenotyping of traits controlling plant water budget," Journal of Experimental Botany, vol. 66, pp. 5581–5593, 9 2015.

[15] Phenospex, "PlantEye F600 multispectral 3D scanner for plants." https://phenospex.com/products/plant-phenotyping/planteye-f600-multispectral-3d-scanner-for-plants/, 2024. Accessed: 2025-01-05.

[16] Z. Ghahramani, Learning dynamic Bayesian networks, pp. 168–197. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998.

[17] Weather Underground, "Weather Hisotry for Patancheru, Telangana, India," 2023.

[18] L. Willocquet, S. Savary, and K. P. Singh, "Revisiting the use of disease index and of disease scores in plant pathology," Indian Phytopathology, vol. 76, p. 909–914, Aug. 2023.

[19] E. González-Domínguez, T. Caffi, V. Rossi, I. Salotti, and G. Fedele, "Plant Disease Models and Forecasting: Changes in Principles and Applications over the Last 50 Years," Phytopathology, vol. 113, pp. 678–693, 5 2023.

[20] A. Kocian, G. Carmassi, F. Cela, L. Incrocci, P. Milazzo, and S. Chessa, "Bayesian Sigmoid-Type Time Series Forecasting with Missing Data for Greenhouse Crops," Sensors 2020, Vol. 20, Page 3246, vol. 20, p. 3246, 6 2020.

[21] M. Baille, A. Baille, and J. C. Laury, "A simplified model for predicting evapotranspiration rate of nine ornamental species vs. climate factors and leaf area," Scientia Horticulturae, vol. 59, pp. 217–232, 11 1994.

[22] C. Stanghellini, Transpiration of greenhouse crops : an aid to climate management. PhD thesis, Wageningen University, Netherlands, June 1987. WU thesis 1152 Proefschrift Wageningen.

[23] Z. Ghahramani and S. T. Roweis, "Learning Nonlinear Dynamical Systems Using an EM Algorithm," Advances in Neural Information Processing Systems, vol. 11, 1998.

[24] A. Kocian, D. Massa, S. Cannazzaro, L. Incrocci, S. Di Lonardo, P. Milazzo, and S. Chessa, "Dynamic Bayesian network for crop growth prediction in greenhouses," Computers and Electronics in Agriculture, vol. 169, p. 105167, 2 2020.

[25] A. Kocian, G. Carmassi, F. Cela, S. Chessa, P. Milazzo, and L. Incrocci, "IoT based dynamic Bayesian prediction of crop evapotranspiration in soilless cultivations," Computers and Electronics in Agriculture, vol. 205, p. 107608, 2 2023.

[26] J. R. L. Filho, A. Kocian, A. A. Frohlich, and S. Chessa, "Continual Learning in Recurrent Neural Networks for the Internet of Things: A Stochastic Approach," Proceedings - IEEE Symposium on Computers and Communications, 2024.

[27] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," Journal of the Royal Statistical Society: Series B (Methodological), vol. 39, pp. 1–22, 12 2018.

[28] C. F. J. Wu, "On the convergence properties of the EM algorithm," The Annals of Statistics, vol. 11, no. 1, pp. 95–103, 1983.

7 Současný stav a plán dalšího výzkumu

V rámci probíhajícího výzkumu jsou řešeny i další navazující oblasti, pro které se připravují výstupy v podobě vědeckých publikací. Jedním z řešených problémů je rozdělení dat pro natrénování neuronové sítě. Datová sada se rozděluje do třech skupin: trénovací, testovací a ověřovací. Na konkrétním rozdělení dostupných dat je závislá úspěšnost natrénování celého modelu. Cílem rozdělení je vytvořit z hlediska komplexnosti homogenní skupiny dat. Z hlediska stratifikovaného rozdělení je nutné najít veličinu, podle které by se data rozdělila. Zaměřil jsem se na komplexitu 3D scény rostlin.

Můžeme najít mnoho přístupů k vyjádření komplexnosti skupiny rostlin nebo jednotlivých rostlin. (Atkins et al., 2023; Hardiman et al., 2011; Hosoi et al., 2013; Parker and Brown).

Přehled metod je uveden v Tabulka 2.

Method	**Metrics**	**Input Data**	**Implementation**
Structural Diversity	FHD, LAI, height profile	LiDAR, RGB-D	Vertical layering
Geometric	FD, SA:V, convex hull ratio	3D point cloud	Shape complexity
Voxel/Grid	Occupancy, entropy, gap fraction	Voxelized 3D point cloud	Density, heterogeneity
Radiative	Gap fraction, LAD, PAR interception	Imaging, LiDAR	Light interception potential
Network/Graph	Branching order, path length	Skeletonized 3D	Architecture, topology
Remote Sensing	Rugosity, CHM stats	LiDAR, UAV, RGB-D	Height & canopy roughness

Tabulka 2 - přehled metod měření komplexnosti rostlin a porostů

Tyto metody vyjadřují složitost například z hlediska propustnosti světla, množství biomasy nebo vertikální hustoty. Metody se většinou zaměřují na vyjádření komplexnosti porostu z hlediska celku. Další metody používané v procesu fenotypizace, které vyjadřují složitost rostlin, popř. skupiny rostlin jsou metody, jež se zaměřují na výšku, šířku a objem nebo na listovou plochu. Většina těchto metod se však zaměřuje na jednotlivé rostliny nebo jejich části. Navíc tyto metody úzce souvisejí s použitými technologiemi nebo využívají pokročilý matematický aparát (Tan et al., 2025; Wang et al., 2021; Wei et al., 2023).

Tyto metody nebyly shledány jako vhodné pro výzkum, jehož cílem je, aby složitost vyjadřovala uspořádání jednotlivých rostlin nebo jejich částí s ohledem na jejich identifikaci. Proto se výzkum zaměřil na stupeň překrytí rostlin. K tomu vedla úvaha, že dvě rostliny rostoucí ze stejného místa se identifikují obtížněji, než dvě samostatné rostliny (Kothawade et al., 2021; Zambrano et al., 2019; Zhang et al., 2021).

Byl vytvořen index, který zohledňuje polohu rostlin mezi sebou a je efektivně algoritmicky vyjádřitelný. Pro tyto účely byl definován index složitosti pro skupinu rostlin (3D sken) v závislosti na jejich překrytí. Index složitosti (angl. Complexity index, dále jen CI), využívá informace z anotovaných dat a je vypočítán jako součet indexu překrytí každého páru rostlin. Index překrytí páru rostlin je poměr průsečíků bází kvádrů, které obsahují každou rostlinu k součtu obsahů obou bází. Pseudokód algoritmu pro výpočet je uveden na Obrázek 2.

Algorithm: Complexity index

Notation: The list of plants in single scan **ψ**; plants pair **ω**; complexity index **CI**;
Input: **ψ**
Output: **CI**

```
2 Initialize CI = 0
3 α = GenerateAllUniquePairs(ψ)
4 for each ω in α
5   CI = CI + OverlapIndex(ω)
6 end for
7 return CI
```

Obrázek 2 – Complexity index algoritmus

Z uvedeného vyplývá že hodnotu CI lze ohraničit vztahem:

$$0 <= CI <= n!/(2!(n-2)!) * 0.5, \quad \text{kde n je počet rostlin}$$

CI byl následně zkoumám z hlediska korelace s jinými indexy vyjadřující komplexitu rostlin. Korelační koeficienty nepotvrdily korelaci s používanými indexy.

Na základě hodnot CI je datová sada rozdělena do příslušných skupin. Výsledky ukazují, že použití CI poskytuje nejlepší hodnoty použitých metrik v porovnání s jinými. V Tabulka 4 je výsledek zvolených metrik při rozdělení dat, podle různých metod rozdělení datových sad. Přehled použitých metod je uveden v Tabulka 3.

Označení metody	Použité proměnné
rnd	náhodné rozdělení
pcl	odrůda; věk
pcl_num	odrůda; věk; počet rostlin
ppl	odrůda ; věk; váženo počtem rostlin
CI-eq	CI, rozdělený do tří skupin podle 3 kvantilů
CI-wei	CI, rozdělený do tří skupin podle 25. a 50. percentilu
CI-eq_species	CI, odrůda, rozdělený do tří skupin podle 3 kvantilů
CI-wei_species	CI, odrůda; rozdělený do tří skupin podle 25. a 50. percentilu
lai_3b	index listové plochy, rozdělený do 3 kvantilů
lai_5b	index listové plochy, rozdělený do 5 kvantilů

Tabulka 3 - přehled proměnných použitých pro rozdělení dat

Výstupy z této části výzkumu jsou součástí plánované publikace č.5.

CI nejen vyjadřuje složitost dat, ale lze jej také použít pro nový typ augmentace, a to modifikaci jednoho skenu podle daného CI. Toho lze dosáhnout buď relativní změnou aktuálního CI, nebo přímým zadáním jeho hodnoty. Výhodou postupu je, že lze měnit index CI kladně i záporně, a tím pozice objektů přiblížit nebo oddálit, a tak zvýšit nebo snížit komplexitu dat. Pro směr pohybu bodu je zvolen vektor spojující střed celého skenu se středem rostliny. Pro pohyb rostlin za účelem dosažení daného CI je vytvořen iterační algoritmus. Ten posouvá rostliny po krocích a po každém pohybu se určí aktuální hodnota CI. Pokud dosažená hodnota dosáhne nebo překročí cílovou hodnotu, algoritmus se zastaví. Cílová hodnota CI musí být nastavena v intervalu od 0 do maximální hodnoty CI. Obrázek 4 – příklad augmentace prostřednictví Complexity indexu. Obrázek 3 představuje pseudokód pro transformaci dat podle daného CI.

Pro augmentační metody v rámci výzkumu bylo navrženo rozdělení metod na realistické a nerealistické. Mezi realistické metody zahrnujeme transformace dat, které zachovávají do vysoké míry realistický vzhled a pozici pozorovaných objektů, v našem případě rostlin. Mezi nerealistické metody řadíme ty, které transformují data do nepřirozených vzhledů a pozic. Smyslem návrhu, implementace a ověřování těchto metod je prozkoumat jejich vliv na výslednou úspěšnost modelu umělé inteligence.

Split method	Shapiro-Wilk p-val	Levene p-val	ANOVA p-val	η^2	ω^2	Variation Coeff.
rnd_0	0,2167	0,0082	0,0002	0,7579	0,6823	10,18%
rnd_1	0,7896	0,2406	0,0139	0,5446	0,4107	5,76%
rnd_2	0,6657	0,5871	0,0000	0,8256	0,7701	9,74%
rnd_3	0,1520	0,4312	0,0010	0,6856	0,5895	8,28%
pcl	0,3312	0,4003	0,0097	0,5682	0,4403	4,61%
pcl_num	0,5541	0,0275	0,0010	0,6855	0,5893	8,40%
ppl	0,4454	0,2254	0,0001	0,7615	0,6870	8,42%
CI-eq	0,8669	0,4432	0,1121	0,3751	0,2002	5,54%
CI-wei	0,9042	0,0150	0,5790	0,1649	-0,0547	5,12%
CI-eq_species	0,5641	0,6629	0,0020	0,6553	0,5507	5,41%
CI-wei_species	0,8186	0,9566	0,0403	0,4667	0,3134	6,86%
lai_3b	0,2301	0,6718	0,0000	0,8358	0,7834	14,94%
lai_5b	0,6173	0,7247	0,0005	0,7186	0,6317	6,39%

Tabulka 4 – hodnoty metrik pro jednotlivé druhy rozdělení

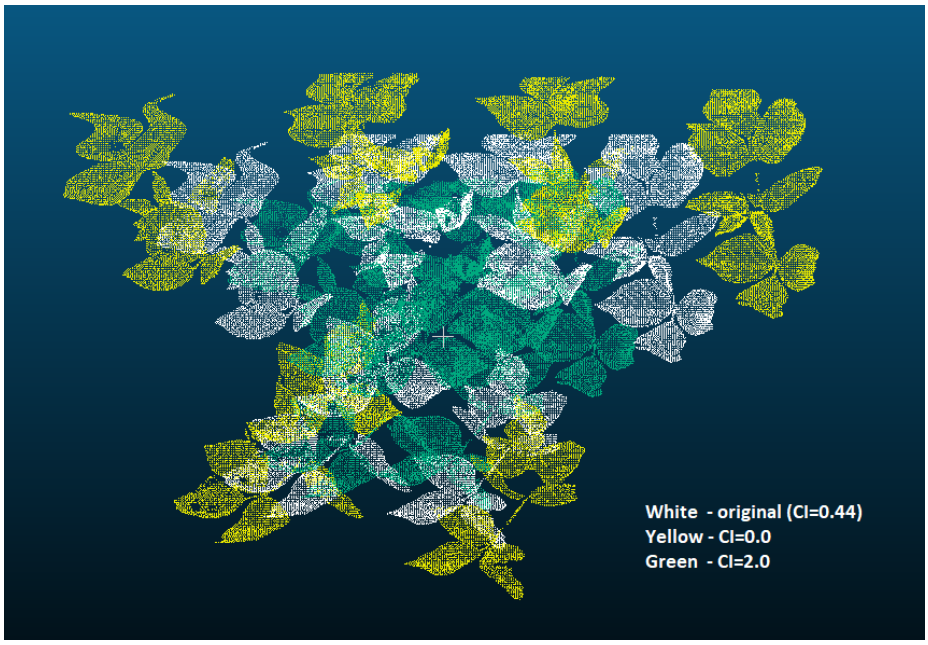
Algorithm: Complexity index transformation

Notation: The list of plants in single scan **ψ**; step value **δ**;
complexity index **CI**; transformated list of plants **ϕ**;
plant **λ**; move vector **γ**; center of scan **σ**;
Input: **ψ, CI, δ**
Output: **ϕ**

```
2 Initialize actualCI = Complexity Index(ψ)
3 Initialize zoomin = (actualCI < CI)
4 Initialize ϕ= ψ
5 Initialiye σ=Center(ψ)
6 while (zoomin and actualCI >= CI) or (not(zoomin) and actual CI <=CI)
7     for each λ  in ϕ
8       γ = Norm(σ – Center(λ))
9       λ = λ + (γ* δ)
10    end for
11    actualCI = Complexity Index(ϕ)
12 end while
13 return ϕ
```

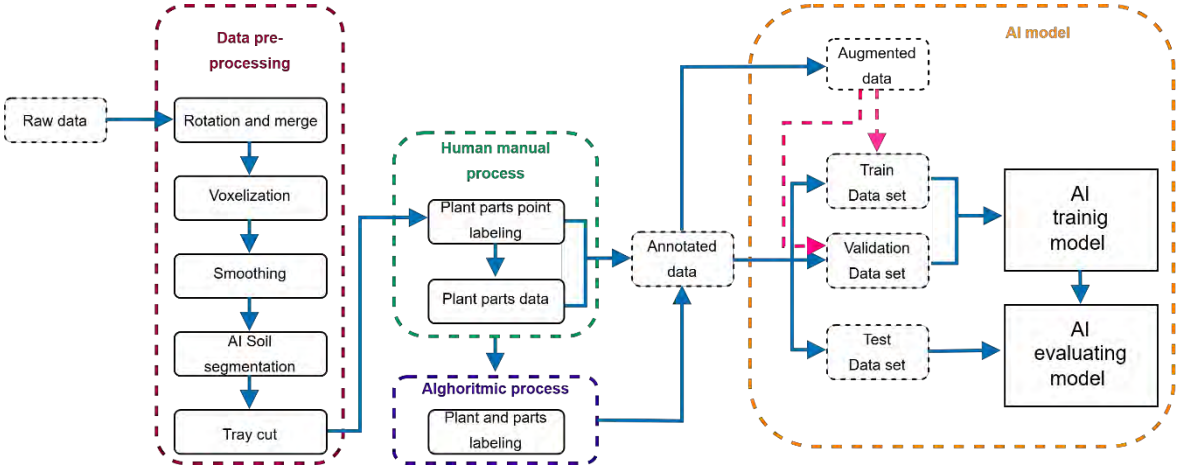
Obrázek 3– algoritmus pro augmentaci pomocí Complexity indexu

Na Obrázek 4 je uveden příklad použití algoritmu využívajícího CI. Bílé rostliny představují originální naskenované rostliny, zelené rostliny transformaci polohy s CI = 2,0 a žluté rostliny s CI = 0,0. Všechny navrhované postupy lze aplikovat nejen na celé rostliny, ale i na jejich orgány. Výstupy z této části výzkumu jsou součástí plánované publikace č.6.



Obrázek 4 – příklad augmentace prostřednictví Complexity indexu

V rámci výzkumu a disertační práce je definován postup pro celý proces získávání dat z vysoce výkonné platformy, jejich zpracování a následnou přípravu pro využití nástrojů umělé inteligence. Tento postup zahrnuje automatizační prvky, jako je použití modelů umělé inteligence pro odstranění pozadí, a také použití algoritmů pro označovací proces. Tyto prvky celý proces urychlují. Schéma implementovaného postupu je znázorněno na Obrázek 5.



Obrázek 5 - schéma postupu zpracování dat z vysoce výkonné platformy

Navržený postup se neomezuje jen na uvedený výzkum, ale je ho možné aplikovat na další oblasti fenotypizačních úloh a na data produkována různými typy zařízení.

8 Závěr

Disertační práce představuje výsledky výzkumu, jež byly úspěšně publikovány ve vědeckých časopisech a na konferencích. Prezentované výsledky potvrdily a ověřily stanovené výzkumné cíle a hypotézy.

Disertační práce se zaměřila na komplexní problematiku přípravy a zpracování dat pro využití nástrojů umělé inteligence ve fenotypizačních úlohách, zejména v prostředí vysoce výkonných metod 3D skenování rostlin rostoucích v zápoji. Cílem bylo navrhnout, implementovat a ověřit nové metodické postupy, které umožní efektivnější, rychlejší a kvalitativně robustnější zpracování fenotypických dat s ohledem na jejich další využití v modelech strojového učení. Všechny výzkumné cíle a hypotézy byly naplněny a experimentálně potvrzeny.

8.1 Odstranění pozadí

V první části práce byla vyřešena zásadní překážka spojená s odstraněním pozadí z 3D skenů. Ukázalo se, že tradiční metody založené na výškové prahové hodnotě či barevných charakteristikách, nedosahují požadované přesnosti, zejména u rostlin v raných vývojových fázích nebo ve scénách s heterogenním pozadím. Proto byl navržen a implementován postup, který využívá neuronové sítě pro automatizovanou segmentaci pozadí. Výsledná metoda významně zvýšila kvalitu očištěných dat a stala se základem dalšího zpracování.

8.2 Datová sada

Ve druhé části práce byla vytvořena a zveřejněna první veřejně dostupná anotovaná datová sada ve formátu 3D bodových mračen, získaných platformou LeasyScan. Datová sada obsahuje 223 multispektrálních 3D skenů několika druhů širokolistých bobovitých rostlin s detailními anotacemi na úrovni orgánů. Tato datová sada vyplňuje významnou mezeru ve vědecké komunitě, neboť dosud neexistovala porovnatelná datová sada vzniklá v reálných venkovních podmínkách a s takto detailní úrovní anotace. Její zveřejnění jako Open Access posiluje transparentnost výzkumu a výrazně podporuje další rozvoj algoritmů pro 3D počítačové vidění rostlin.

8.3 Semi-automatizace

Třetí část práce se zaměřila na proces označování dat, tedy jednu z nejpracnějších a nejnákladnějších fází celého fenotypizačního procesu. Byl navržen semi-automatizovaný postup, který kombinuje manuální anotaci vybraných referenčních bodů a automatizovanou generaci prostorových ohraničení objektů pomocí algoritmu. Současně byl definován systém řízení kvality anotací, který snižuje riziko chyb a zajišťuje konzistenci mezi jednotlivými anotátory, i napříč různými experimentálními sériemi. Tento postup umožnil významně urychlit přípravu trénovacích dat.

8.4 Související výzkum

Čtvrtá část práce demonstruje, že data připravená navrženými metodami lze využít i pro jiné fenotypizační úlohy, například pro predikci průběhu chorob rostlin. V této souvislosti byla aplikována metoda dynamických bayesovských sítí pro modelování vývoje virového onemocnění Sterility Mosaic Disease u plodiny pigeonpea. Výsledky prokázaly, že propojení 3D fenotypických dat a probabilistických modelů, dokáže významně zlepšit přesnost predikce, a tím podporuje šlechtitelské programy minimalizující nutnost manuálních měření.

8.5 Index komplexity

Následně byl v práci navržen a ověřen nový index komplexity, který umožňuje vyjádřit prostorovou složitost rostlinných 3D skenů. Index se ukázal být vhodný jak pro rozdělování datových sad do trénovacích, validačních a testovacích skupin, tak i jako nástroj pro metody augmentace 3D dat. Tato inovace přináší nový přístup k řešení problémů spojených s nerovnoměrnou variabilitou dat a přispívá ke stabilnějšímu a přesnějšímu trénování neuronových sítí.

Disertační práce přináší ucelenou metodiku pro zpracování 3D dat pro použití nástrojů umělé inteligence jako CV v oblasti fenotypizace v celém životním cyklu – od pořízení přes předzpracování, anotace, až po přípravu datových sad a jejich využití v pokročilých analytických úlohách. Navržené metody jsou obecně aplikovatelné i mimo použitou platformu a lze je využít v dalších fenotypizačních, ekologických, agronomických i

průmyslových scénářích. Disertační práce tak přispívá nejen k rozvoji CV v oblasti fenomiky a digitální agronomie, ale i k podpoře využívání umělé inteligence v biologických vědách. Výsledky otevírají nové směry výzkumu a poskytují robustní metodické základy pro budoucí generaci nástrojů pro 3D fenotypování rostlin.

.

Zdroje

Alomar, K., Aysel, H.I., Cai, X., 2023. Data Augmentation in Classification and Segmentation: A Survey and New Strategies. J. Imaging 9, 46. https://doi.org/10.3390/jimaging9020046

Amazon Mechanical Turk [WWW Document], n.d. URL https://www.mturk.com/ (accessed 11.18.25).

Atkins, J.W., Costanza, J., Dahlin, K.M., Dannenberg, M.P., Elmore, A.J., Fitzpatrick, M.C., Hakkenberg, C.R., Hardiman, B.S., Kamoske, A., LaRue, E.A., Silva, C.A., Stovall, A.E.L., Tielens, E.K., 2023. Scale dependency of lidar-derived forest structural diversity. Methods Ecol. Evol. 14, 708–723. https://doi.org/10.1111/2041-210X.14040

Basak, R., Wahid, K.A., 2022. A Rapid, Low-Cost, and High-Precision Multifrequency Electrical Impedance Tomography Data Acquisition System for Plant Phenotyping. Remote Sens. 14, 3214. https://doi.org/10.3390/rs14133214

Behroozpour, B., Sandborn, P.A.M., Wu, M.C., Boser, B.E., 2017. Lidar System Architectures and Circuits. IEEE Commun. Mag. 55, 135–142. https://doi.org/10.1109/MCOM.2017.1700030

Chen, Y., Hu, V.T., Gavves, E., Mensink, T., Mettes, P., Yang, P., Snoek, C.G.M., 2020. PointMixup: Augmentation for Point Clouds. Lect. Notes Comput. Sci. Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinforma. 12348 LNCS, 330–345. https://doi.org/10.1007/978-3-030-58580-8_20/TABLES/4

Cheong, Y., Jun, W., Lee, S., 2024. Performance Enhancement Using Data Augmentation of Point Cloud Based 3D Object Detection for Autonomous Driving, in: 2024 IEEE International Conference on Consumer Electronics (ICCE). Presented at the 2024 IEEE International Conference on Consumer Electronics (ICCE), pp. 1–5. https://doi.org/10.1109/ICCE59016.2024.10444272

Dong, Z., Men, Y., Liu, Z., Li, J., Ji, J., 2020. Application of chlorophyll fluorescence imaging technique in analysis and detection of chilling injury of tomato seedlings. Comput. Electron. Agric. 168, 105109. https://doi.org/10.1016/j.compag.2019.105109

Evers, J.B., van der Werf, W., Stomph, T.J., Bastiaans, L., Anten, N.P.R., 2019. Understanding and optimizing species mixtures using functional–structural plant modelling. J. Exp. Bot. 70, 2381–2388. https://doi.org/10.1093/jxb/ery288

Gavasso-Rita, Y.L., Papalexiou, S.M., Li, Y., Elshorbagy, A., Li, Z., Schuster-Wallace, C., 2024. Crop models and their use in assessing crop production and food security: A review. Food Energy Secur. 13, e503. https://doi.org/10.1002/fes3.503

Ge, Y., Bai, G., Stoerger, V., Schnable, J.C., 2016. Temporal dynamics of maize plant growth, water use, and leaf water content using automated high throughput RGB and hyperspectral imaging. Comput. Electron. Agric. 127, 625–632. https://doi.org/10.1016/j.compag.2016.07.028

Ghanem, M.E., Marrou, H., Sinclair, T.R., 2015. Physiological phenotyping of plants for crop improvement. Trends Plant Sci. 20, 139–144.ICRISAT https://doi.org/10.1016/j.tplants.2014.11.006

Gill, T., Gill, S.K., Saini, D.K., Chopra, Y., de Koff, J.P., Sandhu, K.S., 2022. A Comprehensive Review of High Throughput Phenotyping and Machine Learning for Plant Stress Phenotyping. Phenomics 2, 156–183. https://doi.org/10.1007/s43657-022-00048-z

Guo, Q., Wu, F., Pang, S., Zhao, X., Chen, L., Liu, J., Xue, B., Xu, G., Li, L., Jing, H., Chu, C., 2018. Crop 3D-a LiDAR based platform for 3D high-throughput crop phenotyping. Sci. China Life Sci. 61, 328–339. https://doi.org/10.1007/s11427-017-9056-0

Guo, X., Qiu, Y., Nettleton, D., Schnable, P.S., 2023. High-Throughput Field Plant Phenotyping: A Self-Supervised Sequential CNN Method to Segment Overlapping Plants. Plant Phenomics Wash. DC 5, 0052. https://doi.org/10.34133/plantphenomics.0052

Hahner, M., Dai, D., Liniger, A., Gool, L.V., 2022. Quantifying Data Augmentation for LiDAR based 3D Object Detection. https://doi.org/10.48550/arXiv.2004.01643

Hardiman, B.S., Bohrer, G., Gough, C.M., Vogel, C.S., Curtis, P.S., 2011. The role of canopy structural complexity in wood net primary production of a maturing northern deciduous forest. Ecology 92, 1818–1827. https://doi.org/10.1890/10-2192.1

Hosoi, F., Nakai, Y., Omasa, K., 2013. Voxel tree modeling for estimating leaf area density and woody material volume using 3-D LIDAR data. ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. II-5/W2, 115–120. https://doi.org/10.5194/isprsannals-II-5-W2-115-2013

Kim, S., Lee, S., Hwang, D., Lee, J., Hwang, S.J., Kim, H.J., 2021. Point Cloud Augmentation With Weighted Local Transformations.

Kothawade, G.S., Chandel, A.K., Schrader, M.J., Rathnayake, A.P., Khot, L.R., 2021. High throughput canopy characterization of a commercial apple orchard using aerial RGB imagery, in: 2021 IEEE International Workshop on Metrology for Agriculture and Forestry (MetroAgriFor). Presented at the 2021 IEEE International Workshop on Metrology for Agriculture and Forestry (MetroAgriFor), IEEE, Trento-Bolzano, Italy, pp. 177–181. https://doi.org/10.1109/MetroAgriFor52389.2021.9628564

Kumar, J., Pratap, A., Kumar, S., 2015. Plant Phenomics: An Overview, in: Kumar, J., Pratap, A., Kumar, S. (Eds.), Phenomics in Crop Plants: Trends, Options and Limitations. Springer India, New Delhi, pp. 1–10. https://doi.org/10.1007/978-81-322-2226-2_1

Li, Y., Hu, G., Wang, Y., Hospedales, T., Robertson, N.M., Yang, Y., 2020. Differentiable Automatic Data Augmentation. Lect. Notes Comput. Sci. Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinforma. 12367 LNCS, 580–595. https://doi.org/10.1007/978-3-030-58542-6_35

Liu, Y., Diao, C., Mei, W., Zhang, C., 2024. CropSight: Towards a large-scale operational framework for object-based crop type ground truth retrieval using street view and PlanetScope satellite imagery. ISPRS J. Photogramm. Remote Sens. 216, 66–89. https://doi.org/10.1016/j.isprsjprs.2024.07.025

Ma, J.W., Czerniawski, T., Leite, F., 2020. Semantic segmentation of point clouds of building interiors with deep learning: Augmenting training dataset with synthetic BIM-based point clouds. Autom. Constr. 113, 103144. https://doi.org/10.1016/J.AUTCON.2020.103144

McCormick, R.F., Truong, S.K., Mullet, J.E., 2016. 3D Sorghum Reconstructions from Depth Images Identify QTL Regulating Shoot Architecture. Plant Physiol. 172, 823–834. https://doi.org/10.1104/pp.16.00948

Muhammad, A., Khan, Z.U., Khan, J., Mashori, A.S., Ali, A., Jabeen, N., Han, Z., Li, F., 2025. A comprehensive review of crop stress detection: destructive, non-destructive, and ML-based approaches. Front. Plant Sci. 16. https://doi.org/10.3389/fpls.2025.1638675

Multi-sensor data labeling platform for robotics & AV | Segments.ai [WWW Document], n.d. URL https://segments.ai/ (accessed 4.28.24).

Ndlovu, N., 2020. Application of Genomics and Phenomics in Plant Breeding for Climate Resilience. Asian Plant Res. J. 6, 53–66. https://doi.org/10.9734/APRJ/2020/v6i430137

Nekrasov, A., Schult, J., Litany, O., Leibe, B., Engelmann, F., 2021. Mix3D: Out-of-Context Data Augmentation for 3D Scenes. Proc. - 2021 Int. Conf. 3D Vis. 3DV 2021 116–125. https://doi.org/10.1109/3DV53792.2021.00022

Niemeyer, M., Mescheder, L., Oechsle, M., Geiger, A., 2020. Differentiable Volumetric Rendering: Learning Implicit 3D Representations Without 3D Supervision, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3501–3512. https://doi.org/10.1109/CVPR42600.2020.00356

Paauw, M., Hardeman, G., Taks, N.W., Lambalk, L., Berg, J.A., Pfeilmeier, S., van den Burg, H.A., 2024. ScAnalyzer: an image processing tool to monitor plant disease symptoms and pathogen spread in Arabidopsis thaliana leaves. Plant Methods 20, 80. https://doi.org/10.1186/s13007-024-01213-3

Parker, G.G., Brown, M.J., n.d. Forest Canopy Stratification—Is It Useful?

Patel, A., Lee, W.S., Peres, N.A., Fraisse, C.W., 2021. Strawberry plant wetness detection using computer vision and deep learning. Smart Agric. Technol. 1, 100013. https://doi.org/10.1016/j.atech.2021.100013

Pflüger, T., Jensen, S.M., Liu, F., Rosenqvist, E., 2024. Leaf gas exchange responses to combined heat and drought stress in wheat genotypes with varied stomatal density. Environ. Exp. Bot. 228, 105984. https://doi.org/10.1016/j.envexpbot.2024.105984

PLY - Polygon File Format [WWW Document], n.d. URL https://paulbourke.net/dataformats/ply/ (accessed 11.14.25).

PLY (file format), 2025. . Wikipedia.

Qiu, S., Anwar, S., Barnes, N., 2021. Semantic Segmentation for Real Point Cloud Scenes via Bilateral Augmentation and Adaptive Fusion.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., Tinevez, J.-Y., White, D.J., Hartenstein, V., Eliceiri, K., Tomancak, P., Cardona, A., 2012. Fiji: an open-source platform for biological-image analysis. Nat. Methods 9, 676–682. https://doi.org/10.1038/nmeth.2019

Šebek, P., Pokorný, Š., Vacek, P., Svoboda, T., 2022. Real3D-Aug: Point Cloud Augmentation by Placing Real Objects with Occlusion Handling for 3D Detection and Segmentation. CEUR Workshop Proc. 3349.

Sechidis, K., Tsoumakas, G., Vlahavas, I., 2011. On the Stratification of Multi-label Data, in: Gunopulos, D., Hofmann, T., Malerba, D., Vazirgiannis, M. (Eds.), Machine Learning and Knowledge Discovery in Databases. Springer, Berlin, Heidelberg, pp. 145–158. https://doi.org/10.1007/978-3-642-23808-6_10

Sheshappanavar, S.V., Singh, V.V., Kambhamettu, C., 2021. PatchAugment: Local Neighborhood Augmentation in Point Cloud Classification.

Shi, P., Qi, H., Liu, Z., Yang, A., 2023. Context-guided ground truth sampling for multi-modality data augmentation in autonomous driving. IET Intell. Transp. Syst. 17, 463–473. https://doi.org/10.1049/itr2.12272

Shiflet, A.B., Shiflet, G.W., 2014. Introduction to Computational Science: Modeling and Simulation for the Sciences - Second Edition. Princeton University Press.

Supervisely: Curate, Label and Build Production Models in One Platform [WWW Document], n.d. URL https://supervisely.com/ (accessed 11.18.25).

Tan, Y., Li, Y., Jia, S., Zhao, Q., 2025. Improving coniferous forests leaf area index estimation by filling the occluded point cloud from airborne laser scanning. Measurement 242, 115866. https://doi.org/10.1016/j.measurement.2024.115866

Vadez, V., Kholová, J., Hummel, G., Zhokhavets, U., Gupta, S.K., Hash, C.T., 2015. LeasyScan: a novel concept combining 3D imaging and lysimetry for high-throughput phenotyping of traits controlling plant water budget. J. Exp. Bot. 66, 5581–5593. https://doi.org/10.1093/jxb/erv251

Wang, C., Ma, C., Zhu, M., Yang, X., Key, M., 2021. PointAugmenting: Cross-Modal Augmentation for 3D Object Detection.

Wang, Q., Tan, Y., Mei, Z., 2020. Computational Methods of Acquisition and Processing of 3D Point Cloud Data for Construction Applications. Arch. Comput. Methods Eng. 27, 479–499. https://doi.org/10.1007/s11831-019-09320-4

Wei, B., Ma, X., Guan, H., Yu, M., Yang, C., He, H., Wang, F., Shen, P., 2023. Dynamic simulation of leaf area index for the soybean canopy based on 3D reconstruction. Ecol. Inform. 75, 102070. https://doi.org/10.1016/j.ecoinf.2023.102070

Weon, I.-S., Lee, S.-G., Ryu, J.-K., 2020. Object Recognition Based Interpolation With 3D LIDAR and Vision for Autonomous Driving of an Intelligent Vehicle. IEEE Access 8, 65599–65608. https://doi.org/10.1109/ACCESS.2020.2982681

Wieringa, R.J., 2014. Design Science Methodology for Information Systems and Software Engineering. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-43839-8

Wu, M., Huang, H., Fang, Y., 2022. 3D Point Cloud Completion with Geometric-Aware Adversarial Augmentation. Proc. - Int. Conf. Pattern Recognit. 2022-August, 4001–4007. https://doi.org/10.1109/ICPR56361.2022.9956045

Xiong, X., Yu, L., Yang, W., Liu, M., Jiang, N., Wu, D., Chen, G., Xiong, L., Liu, K., Liu, Q., 2017. A high-throughput stereo-imaging system for quantifying rape leaf traits during the seedling stage. Plant Methods 13, 7. https://doi.org/10.1186/s13007-017-0157-7

Zambrano, J., Fagan, W.F., Worthy, S.J., Thompson, J., Uriarte, M., Zimmerman, J.K., Umaña, M.N., Swenson, N.G., 2019. Tree crown overlap improves predictions of the functional neighbourhood effects on tree survival and growth. J. Ecol. 107, 887–900. https://doi.org/10.1111/1365-2745.13075

Zhang, W., Xu, X., Liu, F., Zhang, L., Foo, C.S., 2021. On Automatic Data Augmentation for 3D Point Cloud Classification. 32nd Br. Mach. Vis. Conf. BMVC 2021.

Zheng, Y., Zhang, Z., Yan, S., Zhang, M., 2021. Deep AutoAugment. Presented at the International Conference on Learning Representations.

Zhou, J., Applegate, C., Alonso, A.D., Reynolds, D., Orford, S., Mackiewicz, M., Griffiths, S., Penfield, S., Pullen, N., 2017. Leaf-GP: An Open and Automated Software Application for Measuring Growth Phenotypes for Arabidopsis and Wheat. https://doi.org/10.1101/180083

Zhou, X., Luo, X., 2009. Advances in Non-Destructive Measurement and 3D Visualization Methods for Plant Root Based on Machine Vision, in: 2009 2nd International Conference on Biomedical Engineering and Informatics. Presented at the 2009 2nd International Conference on Biomedical Engineering and Informatics, pp. 1–5. https://doi.org/10.1109/BMEI.2009.5304876

Zhu, Q., Fan, L., Weng, N., 2024. Advancements in point cloud data augmentation for deep learning: A survey. Pattern Recognit. 153, 110532. https://doi.org/10.1016/J.PATCOG.2024.110532

Zini, S., Gomez-Villa, A., Buzzelli, M., Twardowski, B., Bagdanov, A.D., Weijer, J. van de, 2023. Planckian Jitter: countering the color-crippling effects of color jitter on self-supervised training. https://doi.org/10.48550/arXiv.2202.07993

9 Příloha - Přehled publikací autora

9.1 Článek impaktovaný

GALBA, Alexander; MASNER, Jan; KHOLOVÁ, Jana; KARTAL, Serkan; STOČES, Michal et al. Annotated 3D Point Cloud Dataset of Broad-Leaf Legumes Captured by High-Throughput Phenotyping Platform. Online. *Scientific Data*. 2025, vol. 12, no. 1. ISSN 2052-4463. Dostupné z: https://doi.org/10.1038/s41597-025-06049-7. [cit. 2025-11-10].

KARTAL, Serkan; MASNER, Jan; KHOLOVÁ, Jana; **GALBA, Alexander**; MURUGESAN, Tharanya et al. AI-Driven Background Segmentation for High-Throughput 3D Plant Scans. Online. *IEEE Access*. 2025, roč. 13, s. 136027-136037. ISSN 2169-3536. Dostupné z: https://doi.org/10.1109/access.2025.3594406. [cit. 2025-11-08].

GALBA, Alexander; KÁNSKÁ, Eva; MIKEŠ, Vojtěch; VANĚK, Jiří a JAROLÍMEK, Jan. Application of Quality Management System in the Research Process: A Case Study for Plant Phenotyping Research. Online. *Agris on-line Papers in Economics and Informatics*. 2024, roč. 16, č. 4, s. 79-86. ISSN 1804-1930. Dostupné z: https://doi.org/10.7160/aol.2024.160406. [cit. 2025-11-19].

KUBATA, Karel; **GALBA, Alexander**; ŠILEROVÁ, Edita; OČENÁŠEK, Vladimír a CIHELKA, Petr. Practical use of Agriculture 4.0 digital technologies to meet the EU's strategic goals in Czech agriculture. Online. *Agris on-line Papers in Economics and Informatics*. 2024, roč. 16, č. 2, s. 63-74. ISSN 1804-1930. Dostupné z: https://doi.org/10.7160/aol.2024.160205. [cit. 2025-11-19].

9.2 Stať ve sborníku

MIKEŠ, Vojtěch; KOCIAN, Alexander; KHOLOVÁ, Jana; MASNER, Jan; KLECZKOWSKI, Adam; SHARMA, Mamta; CHESSA, Stefano**; GALBA, Alexander;** ŠIMEK, Pavel. Forecasting Sterility Mosaic Disease in Pigeonpea Using Dynamic Bayesian Networks and 3D Point Cloud High-throughput Scanning Platform. Online. In: *2025 21st International Conference on Intelligent Environments (IE)*. IEEE, 2025, s. 1-8. Dostupné z: https://doi.org/10.1109/ie64880.2025.11130066. [cit. 2025-11-10].

KOVÁŘ Lukáš; STOČES Michal; KUBATA Karel; JAROLÍMEK Jan; **GALBA Alexander**; HAVRÁNEK Martin a NOVÁK Vojtěch. CULS – INDOOR OCCUPANCY DETECTION DATASET . In: Agrarian perspectives XXXII. Human Capital and Education in Agriculture. Praha:PEF ČZU v Praze , 2023, s. 143-155, 978-80-213-3309-3.