Title

A primer on artificial intelligence in plant digital phenomics: embarking on the data to insights journey

Journal
Trends in Plant Science, 28(2)

ISSN
1360-1385

Authors
Harfouche, Antoine L
Nakhle, Farid
Harfouche, Antoine H
et al.

Publication Date
2023-02-01

DOI
10.1016/j.tplants.2022.08.021

A primer on artificial intelligence in plant digital phenomics: embarking on the data to insights journey

Antoine L. Harfouche,[1,*] Farid Nakhle,[1] Antoine H. Harfouche,[2] Orlando G. Sardella,[1] Eli Dart,[3] and Daniel Jacobson[4]

[1]Department for Innovation in Biological, Agro-food and Forest systems, University of Tuscia, Viterbo, VT 01100, Italy

[2]Unité de Formation et de Recherche en Sciences Économiques, Gestion, Mathématiques et Informatique, Université Paris Nanterre, 92001 Nanterre, France

[3]ESnet, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

[4]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA

*Correspondence: aharfouche@unitus.it (A. L. Harfouche)

Keywords

AI system architecture, black box models, data analytics, digital phenomics, explainable artificial intelligence, interpretable by design models

Artificial intelligence (AI) has emerged as a fundamental component of global agricultural research poised to impact many aspects of plant science. In digital phenomics, AI is capable of learning intricate structure and patterns in large datasets. Here, we provide a perspective and primer on AI applications to phenome research. We propose a novel human-centric explainable AI (X-AI) system architecture, consisting of data architecture, technology infrastructure, and AI architecture design. We clarify the difference between post-hoc models and interpretable by design models. We include guidance for effectively using an interpretable by design model in phenomics analysis. We also provide a direction to sources of tools and resources for making data analytics increasingly accessible. This primer is accompanied by an interactive online tutorial.

Approaching plant phenomics from different angles

Crop breeding relies heavily on phenotypic information, which remains a bottleneck for realizing its full potential. The advent of plant phenomics (topic reviewed in [1–9]), which broadly can be considered the systematic study of phenotypes, however, marks a turning point. Phenomics platforms equipped with novel imaging sensors promise to make it possible to perform phenotyping of a wide range of plant traits, organs, and environmental situations at scale (Figure 1). These new technological developments have opened up avenues for automated data acquisition, evolving phenomics to become a thriving research field of its own [10].

Embracing digital technology holds tremendous potential for driving transformative changes in plant phenomics by improving the collection of, access to, and analysis of phenomic big data. As such, digital phenomics furnishes tools and resources that aid in the digitization of plant phenomics. It uses phenome data and metadata to guide decision-making along the entire data analytics cycle [11]. As data grows in volume and varies in sources, effective management strategies must be put into place (Figure 2). Digital phenomics gives rise to computational phenomics, which allows the assembly of a broad array of methods that aid in the discovery of intricate structure and patterns from phenotypic data, using technology infrastructure (Figure 3) and artificial intelligence (AI) architecture design (Figure 4).

In the past decade, AI – the science of studying, designing, and developing intelligent computer systems that can perform tasks that normally require human intelligence – finally began to reveal its remarkable power and disruptive potential. Driven mainly by the advent of machine learning (ML) – a particular approach to AI in which intelligent systems learn and derive models from training datasets – and deep learning (DL) – a specialized branch of ML that leverages neural networks to spot patterns in complex data – AI flexed its muscles by achieving predictive successes in phenomics. For example, in red–green–blue (RGB) image analysis, convolutional neural networks (CNNs) were used to predict the yield of individual plants of barley and wheat [12], to classify and quantify biotic and abiotic stresses in leaves of various fruit and vegetable crops [13], to segment roots of chicory, wheat, and rapeseed [14–16], and to count tobacco leaves [17]. A CNN was also used to provide explainable classifications of biotic and abiotic stresses in soybean leaves by isolating the top-k feature maps learned by the model [18]. A random forest (RF), a neural network (NN), a k-nearest neighbor (KNN), a partial least squares (PLS), and a support vector machine (SVM) were employed to estimate nitrogen nutrition index for improving nitrogen use efficiency in rice [19,20]. In wood anatomical images at a microscopic level, a mask region-based CNN (Mask R-CNN) was used to analyze the intrinsic variability of wood anatomical features in conifers, alder, beech, and oak [21]. In analyzing multispectral and hyperspectral data, a CNN was used to provide explainable identification of biotic stress in individual soybean plants by incorporating saliency maps [22]; a PLS and a RF to estimate above ground biomass in maize [23]; an SVM, a KNN, and a linear discriminant analysis (LDA) to detect and segment root decay in wheat [24]; an LDA and a PLS to detect response to drought stress in bell pepper, courgette, sunflower, radish, foxtail millet, and sorghum [25]; and a data mining sharpener to guarantee consistent spatial resolution among heterogeneous remote sensing image datasets to dissect the latent heat flux signature of poplar in response to drought [26]. In three-dimensional (3D) point cloud analysis, an SVM was applied to estimate yield and canopy geometric characterization in apple [27]. In thermal infrared (TIR) image analysis, SVMs and Gaussian processes were used to identify drought stress in spinach [28] and rotation forests were used to predict plant water status in grapevine [29]. In chlorophyll fluorescence image analysis, a CNN was used to identify abnormalities in organelle morphology in Arabidopsis [30]. In X-ray computed tomography (X-ray CT) analysis, an encoder-decoder network was used to segment wheat roots [31].

Analyzing data coming from different sensors and imaging techniques of the same biological sample (i.e., plant) simultaneously can further improve phenotypic trait predictions. Recent studies demonstrated that fusion of multiple data sources originating from the same plants (i.e., paired data) perform better than a single source.

For example, using a deep NN (DNN), the fusion of RGB, TIR, and multispectral data delivered superior performance over single sensor data analytics for yield prediction in soybean [32]; similar results were reported for the estimation of soybean chlorophyll content, nitrogen concentration, leaf area index, and above ground biomass by employing an extreme learning machine [33]. Now and in the future, rigorous data integration of phenomics and other different omics datasets that were not originally set out to be integrated and are of distinct biological samples (i.e., unpaired data) may help dissecting biological mechanisms that underlie desirable traits and shed light on the flow of information that underpins plant responses to environmental stresses [10,34–37]. And because the phenotype of a plant is the result of interaction between its genotype and the environment (G x E) in which it grows [38], integration efforts should also include environmental data such as climatypes; this will be crucial for designing new crop ideotypes that are optimized for niche environments in a world with a rapidly changing climate [34].

Importantly, data management strategies should incorporate the findable, accessible, interoperable, and reusable (FAIR) guiding principles [39] to put those phenome and envirome data to their most effective use. This requires standards to ensure that necessary metadata are recorded about data generation methods and the experimental and environmental conditions in which they were acquired [40]. In this regard, the minimum information about a plant phenotyping experiment (MIAPPE) standard has been a great step forward to harmonize data from phenotyping experiments with controlled vocabulary and ontologies [41,42]. Accordingly, the development of tools for capturing the complete set of metadata is poised to have high impact on the support of

FAIR data. As standards and tools become more widely disseminated and explored, we envisage metadata becoming commonly annotated by users, expected by referees, and required by journals and data repositories. Furthermore, the workflows that are used to analyze data should themselves be FAIR [43]. When data sharing is not viable due to possible privacy or security concerns, **federated learning** (FL; see Glossary) and communication-efficient FL offer an unprecedented opportunity to train AI models without sharing data [44,45]. FL gained traction in medical imaging applications [46–48] and carries great promise for overcoming data sharing challenges in plant phenomics.

This primer provides suggestions on how to use AI effectively in plant phenomics, on how to ensure that human-centric explainable AI (X-AI) can benefit all, and discusses various X-AI approaches and techniques. We have created a central directory of all publicly available plant imaging datasets, and report their sources, accessibility, and a summary of species and organ systems represented (Table 1). This review is accompanied by an interactive tutorial to train an interpretable by design model to deliver predictive and prescriptive analytics to users. Our primer is intended as an educational resource for phenomicists who are interested in applying X-AI approaches and techniques, and plant scientists who seek a high-level understanding of this rapidly evolving field. Data scientists and information systems (IS) scientists may also use this primer as an introduction to the promising applications of X-AI in phenomics.

How to use AI effectively: ménage-à-trois between plant science, data science, and IS

Plant science has the potential to provide innovative solutions for the world's most pressing challenges; however, recent advances in discovery methods have greatly accelerated our ability to collect data, leaving us with the challenge of analyzing, interpreting, and integrating the plethora of data [49]. To handle such data, data science has attracted a lot of attention, promising to turn data into useful predictions and insights [50]. To do that, data science needs supporting resources including algorithms, software, and hardware infrastructure. IS combines those resources to create AI architecture designs, and to transform, store, and distribute data for analysis. While the relationship between these disciplines has not been reinforced repeatedly in history, today with the depth of data analysis, the scale and dimension of the data, and the nature of the scientific questions, an interaction in a ménage-à-trois fashion is highly needed.

AI and the bias cascade

Multiple sources of **bias** can affect the performance of AI systems used in phenomics and can occur across the different development steps of AI applications: data collection or selection, data preprocessing, model development, model evaluation, and deployment. Introduced bias can have a domino effect as it propagates from its entry point to the succeeding development steps, creating a bias cascade.

The bias cascade starts with the data collection or selection step, where experimental data are collected or selected from publicly available datasets (Table 1). Here, bias can occur for a number of reasons: (i) 'measurement' bias, when data contains faulty measurements originating from instrumentation malfunctions, wrong values from miscalibrated sensors, or errors of precision that result in data distortion [51]; (ii) 'label' bias, when data is laden with subjective judgments of human experts and thus inconsistently or wrongly labeled [52]; (iii) 'sample selection' bias, when the training data does not represent a random sample from the entire dataset [53], causing a model to ignore data belonging to classes that were not represented during the data selection process; and (iv) 'group attribution' bias, when a data sample is selected from an incorrect target population [53], where a model can fail to distinguish between some classes and consider them the same.

Data preprocessing is performed to eliminate noisy (e.g., blurred images, images with unfavorable lighting conditions, images that do not represent the object of interest), incomplete (e.g., unannotated images), duplicate data, and to normalize datasets as needed to account for batch effects (e.g., groups of images taken under different lighting conditions or with different camera settings) or systematic experimental artifacts (e.g., reflections in images). In this step, even if the training set was representative of the entire dataset, data can be intrinsically **unbalanced** where certain plant species, genotypes, or even stresses are underrepresented. Such cases can introduce the 'class imbalance' bias.

Bias may also arise during the model development and evaluation steps, where a model is trained and its ability to generalize beyond the training set, on new, previously unseen data is evaluated. As most AI algorithms identify correlations between variables in the underlying data but without being able to detect causal relations, two biases are likely to arise: (i) the 'correlation fallacy' that confuses correlation with causation [53] where a model wrongly deduces a cause-and-effect relationship between correlated variables; and (ii) the ‘apophenia’ when a model sees patterns while none actually exist [54]. These two biases can be amplified when a massive quantity of training data is used, mistakenly offering connections that radiate in all directions [54], and producing probable yet uncertain predictions. Further, training complex models (i.e., models with many trainable parameters) can capture noise-generated patterns, tricking them into thinking that the noise encodes real information [55]. This problem introduces the ‘overfitting’ bias and causes a steep drop-off in predictive performance at the evaluation step. Similarly, such performance drop-offs also occur when models are unable to accurately capture relationships between variables and thus introducing the ‘underfitting’ bias [56].

Finally, at the deployment step, bias can occur in situations where data used in practice differs from training data (e.g., different weed or crop species), which is known as the ‘domain shift’ bias.

Creating a human-centric X-AI

It is therefore crucial to mitigate bias to increase the success probability of the AI algorithm for the task at hand. Let alone that bias mitigation serves as a building block towards AI **trustworthiness** [57]. So, what can be done to mitigate detrimental biases in AI in plant phenomics? There is a consensus on the need to develop a human-centric X-AI system that will not just aspire to meet human requirements regarding explainability and trustworthiness, but, more importantly, will actively aim to keep a **human-in-the-loop** (HITL) for a harmonious human and AI system symbiosis. We believe that such a system should not only put humans at its center, but also integrate their knowledge into its predictive process.

Designing human-centric X-AI for plant phenomics is not without challenges; it requires a dedicated and multidisciplinary team effort, involving plant scientists, data scientists, and IS scientists to bring AI to its most feasible, desirable, viable, and responsible state. This novel multidisciplinary knowledge is clearly imperative to identify and reduce AI biases, and to facilitate explainability and accountability. We advocate that such a system architecture is required to constantly realign data architecture and technology infrastructure to serve novel AI architecture designs. Phenotyping complex traits demands the integration of data on different morphological, physiological, temporal, geospatial, and environmental variables [35,58]. While large datasets are vital for creating accurate AI models and validating their results, storing them in a FAIR manner can be challenging. Data architecture plays a fundamental role in meeting these requirements. It consists of a set of standards that govern which data is collected, whether it should be transformed (e.g., data cleaning, deduplication, format conversion, structuring, validation, etc.) before or after storage using extract, transform, load (ETL) or extract, load, transform (ELT) processes, and where (data warehouses or data lakes) and how (matrices, cubes, polytopes, or distributed in-memory) it is stored (Figure 2). Without AI, these data streams would be overwhelming and chaotic [35], but reaching the full potential of AI-based analysis of large phenomic datasets comes down to the right technology infrastructure which defines the components that serve as a foundation for the data life cycle, including hardware infrastructure, network flow, software frameworks, and programming languages (Figure 3). High performance computing

(HPC), like pre-exascale supercomputers, is boosting both the accuracy and predictive power of these approaches. While central processing units (CPUs) maximize the performance of an algorithm, graphics processing units (GPUs) can dramatically increase AI training speed thanks to their processing cores initially designed to process visual data such as images and videos [11]. For example, to take advantage of GPUs, the compute unified device architecture (CUDA) software framework provides a development environment for creating and optimizing AI applications on **GPU-accelerated** local computers or supercomputers. However, CUDA works exclusively on Nvidia GPUs; alternatively, the open computing language (OpenCL) and openACC frameworks work on multiple types of GPUs [59]. Another option is to translate automatically CUDA source code into portable heterogeneous-computing interface for portability (HIP) using source-to-source translators such as HIPify, so that non-Nvidia GPUs can benefit from the rapid development of CUDA applications. Additionally, software libraries such as kokkos, RAJA, open multi-processing (OpenMP), and one application programming interface (oneAPI) can be leveraged to unlock the promise of heterogeneous computing where **compute nodes** employ more than one type of processors including CPUs, GPUs, and tensor processing units (TPUs), among others. This enables the development of scalable AI-based applications in a hardware agnostic way. With the advent of exascale computing, supercomputers will deliver higher performance in pattern searching in phenomic big data, and thus, will boost AI abilities in digital phenomics, speeding up crop design (Figure 3A). But, building powerful supercomputers is a never-ending race, and as new ones get launched, the number of compute nodes they comprise increases. For example, the first supercomputer to break the exascale barrier, Summit, comprises 4,608 compute nodes, while the most powerful exascale supercomputer that tops the latest TOP500 list[i], Frontier, contains 9,472. This makes it harder to exploit supercomputers efficiently because of their need to transmit data back and forth between their nodes, running huge numbers of computations at the same time [60]. Implementing AI algorithms (Figure 4A) for such **parallel computing** is not easy. Luckily, emerging free and open-source software frameworks such as Tensorflow Keras, PyTorch, scikit-learn, and XGBoost, among others, and software libraries such as cuNumeric are enabling scalability on parallel computing. As more powerful exascale supercomputers are being anticipated [61], researchers may start to utilize quantum computers at some point in the future [62]. This will ultimately drive digital phenomics towards designing faster, better crops and providing sustainability-friendly solutions (Figure 3A). Beside the hardware infrastructure, properly designed network flows (Figure 3B), such as the ‘science demilitarized zone (DMZ)' that includes network architecture and performance tools [63], enable high-throughput access to datasets in a secure and timely manner while conforming with the FAIR data principles [39]. Software frameworks (Figure 3C) provide a working environment that helps researchers achieve higher productivity in designing AI algorithms; they support more than one programming language (Figure 3D), enabling fast and efficient implementation of algorithms without compromising code quality.

Because data are only as good as the tools and algorithms available to analyze them, solving complex biological questions requires a creative process during which efficient AI algorithm architectures are designed and developed. Customized algorithms and architectures can leverage currently available AI architecture designs (Figure 4) to come up with new architecture designs tailor-made to find answers to the questions at hand. Such promising designs should combine knowledge-based AI, to represent human expert knowledge, with data-driven AI to discover connections and correlations automatically in big data. This combination will result in an informed AI that acquires both **tacit** and **explicit knowledge** of its designers (e.g., the interaction between data scientists, IS scientists, and plant scientists) and users (e.g., breeders and farmers), and integrates that tacit and explicit knowledge with knowledge discovered from data and metadata.

It is noteworthy, however, that new architecture designs should also integrate knowledge into X-AI to enable the monitoring of the inputs and outputs of the algorithms, provide more human-comprehensible explanations for their decisions, deliver superior performance, mitigate bias, and aid in verifying models' adherence to ethical and socio-legal values. Ensemble methods can, for example, be leveraged to design new AI algorithms that are both informed and explainable (Figure 4B). Ultimately, improvements in informed X-AI would help develop novel interpretable algorithms and are likely to be crucial to enable human-centric X-AI in phenomics.

Mitigating bias in human-centric X-AI

A human-centric X-AI system is emerging, whereby plant scientists, data scientists, and IS scientists must work together to seize this opportunity to help identify and mitigate bias by using a number of strategies.

Starting from the top of the bias cascade, at the data collection or selection step, the minority of data that do not conform to the general characteristics of a given dataset, known as outliers, should be removed during data cleaning to mitigate the 'measurement' bias. As for the 'label' bias, data annotators should be supplied with detailed instructions containing visual examples of the correct output for a given input to be able to reduce ambiguities and avoid mistakes that result from incorrect or incomplete knowledge. For example, when labeling weed species, in addition to their morphological descriptions, a visual representation of each species could be helpful for annotators. Next, the 'sample selection' and 'group attribution' biases can be mitigated by establishing random sample selection and statistical correction processes [64,65].

When preprocessing data, intrinsically unbalanced datasets can be balanced by means of oversampling (i.e., augmenting the number of training examples within the minority class to be equivalent to other classes) and/or undersampling (i.e., reducing the number of training examples within the majority class to be equivalent to other classes) [11] to eliminate the 'class imbalance' bias [66].

Properly sampled and preprocessed data mitigate the risk of 'correlation fallacy' and 'apophenia' biases, which can occur during the model development step. When training, 'overfitting' can be debiased by either increasing the size of training data, decreasing the model complexity, or ignoring the less important features in a process called regularization [67]. Whereas 'underfitting' bias can be resolved by increasing the complexity of the model to capture nonlinear relationships in data. During the evaluation step, models yielding incorrect predictions such as misclassifying crops as weeds or vice versa should be inspected carefully; X-AI can be leveraged to better understand how the model reached its predictions which helps identify previously unknown bias. However, post-hoc approaches to explainability are not necessarily transparent (i.e., because they only approximate models' prediction procedure), and thus, it might be better to employ interpretable by design models (see next section).

Notably, to avoid the 'domain shift' bias and identify unwanted biases in deployment, it is crucial that the model is carefully monitored to assess whether the data being used in practice are representative of those used during training. It is important to note that risk-based regulations of AI are on the horizon in the US[ii] and Europe[iii,iv]. When new regulations enter into force, post-authorization monitoring of AI applications becomes crucial to ensure that the performance of models does not degrade in practice. Once a model has passed regulatory authorization and is implemented in phenomics, it needs to be retrained periodically using new datasets to prevent it from becoming outdated, ensuring 'domain shift' bias mitigation.

Reducing the risk of bias in AI models requires continuous human attention across the five development steps, keeping HITL. Studies have shown that human-computer interaction in HITL AI has improved the predictive performance of AI-based image analysis and reduced biases [68–70]. HITL can make a significant impact in phenomic data collection, data preprocessing, model development, evaluation, and deployment. It plays a critical role in the collection and preparation of data to be used for training an AI model. As such, model training is often a HITL iterative process that identifies biases or weaknesses of the model (e.g., images on which the model fails due to incomplete training sets or inappropriate parameterization) and adjusts the training set and parameters to reduce any biases and ensure the best model performance. It is recommended to start each training step with small iterations and plan on how the feedback of the team of humans can be collected and propagated to other steps, relying on their intelligence to perform complex tasks. This paradigm allows leveraging the advantages of AI while having humans at various checkpoints to fill gaps where models are not confident in their predictions or where they may fall short due to underlying biases [71]. HITL may also offer advantages to evaluating the accuracy of AI predictions and interpreting their decisions by interacting with explainable models. The benefits of HITL extend to deployment by monitoring the model for possible biases and ensuring the reliability of the AI system. HITL can feedback into itself to respond to changes in the real-world environment. For example, after data collection and preprocessing, in each training iteration, plant scientists are shown a list of misclassified images with the outputs of the AI algorithm to hand-verify predictions and assess false positives and false negatives. For instance, the model might misclassify crops as weed; but this could be due to an algorithmic or learning error, or to mislabeled images. They then correct the wrong labels, if any, to ensure high-quality data. Data scientists evaluate the model, tune its **hyperparameters**, and retrain it. Such iterations between humans and AI are effective to generate training data based on human judgment to increase learning efficiency and enhance model performance [71]. Data scientists can provide the expertise necessary to help IS scientists design AI architectures with explanatory capacity supported by theoretical underpinnings. Finally, HITL monitors the model outcomes post deployment to ensure that all biases are identified and mitigated. Furthermore, ensembles of models can be used in the HITL process. An intrinsically interpretable model such as an iterative RF [72,73], can be used initially for feature engineering to determine the variables (e.g., wavelet decomposition in RGB and hyperspectral images) that will then be used in a deep learning predictive model.

This three-way collaboration can amplify knowledge about domain-specific feature engineering and selection to reach a level of augmented intelligence that can help discovering new ways to make AI more efficient, less biased, and explainable. It also creates new opportunities for human-centric X-AI to predict desirable phenotypic traits and aid efforts to breed climate-proof crops fast enough.

How to move from data inputs to outcomes: opening the black box or designing a transparent glass box for explainability

AI continues to permeate plant phenomics as recently reviewed in [74–76]. However, complex AI models are difficult to explain even among data scientists; they operate as black boxes and require a leap of faith to believe their predictions [35]. Explainability of AI models would not only increase the trust of users in why and how predictions were made but also help data scientists enable better diagnostics and enhance their performance. Although these desirable properties of explainability have led to a recent growing interest in X-AI research [77], its origin traces back to the early 1970s when Edward Shortliffe introduced the AI-based antimicrobial therapy consultation system for assisting physicians who need advice about appropriate therapy. The system made use of a set of decision rules coded, categorized, and hand-entered into it to give advice and explain the reasons behind its predictions [78]. In 1979, Jon Doyle introduced the truth maintenance systems (TMS), an independent module that constructs explanations of predictions by recording and maintaining a representation of the knowledge acquired by an expert system [79]. TMS research and development continued until the 1990s, when researchers began to study the possibility of extracting meaningful explanations from non-hand-coded rules that are generated by trained models such as NN [80].

The rise of DL in the 2010s [11] increased the complexity of AI models and consequently, the demand for X-AI algorithms. To address this issue, researchers have been developing new approaches and techniques to make these models explainable. Unfortunately, the rush in X-AI development has caused confusion on its various approaches in the literature, where they are not accurately described and are often confused together [81]. While all those approaches revolve around allowing humans to observe how predictions of an AI model came to be, we can technically distinguish between research involving post-hoc models and interpretable by design models.

As current AI models are often developed with only predictive performance in mind, post-hoc algorithms can be used to explain them. They are employed after a black box model is trained and are not connected to its internal design; they can either be model-specific or model-agnostic [82]. In principle, model-specific algorithms are limited to certain black box models. For example, DL important features (DeepLift) is a model-specific algorithm that can explain DNNs and does not work for any other algorithm. On the other hand, model-agnostic algorithms such as the local interpretable model-agnostic explanations (LIME) [83] and Shapley additive explanation (SHAP) [84] are more general and can be applied to any black box model. Commonly, post-hoc algorithms work by: (i) probing or inspecting the trained parameters to understand what has the black box model learned; (ii) employing data perturbation strategies which involve modifying the input data and observing the changes in the black box model predictions; or (iii) using a more interpretable model (e.g., decision tree) referred to as a surrogate model to approximate and provide explanations of predictions made by the black box model. Recently, researchers started applying post-hoc algorithms in plant phenomics to identify, classify, and quantify plant stresses [11,85–89] and to count leaves [90]. However, as post-hoc algorithms approximate the inner workings of black box models, it is possible that their generated explanations do not provide enough detail to understand what the black box model is actually doing [91]. On the other hand, interpretable by design algorithms do not need an additional (post-hoc) algorithm to be explainable; they provide their own explanations, which are faithful to what the model actually computes [91].

These algorithms have existed since the development of expert systems in the 1970s. They have, however, been labeled as less accurate because scientists argue that there is a tradeoff between accuracy and explainability in a way that, the highest performing algorithms are the least explainable, and the most explainable ones are less accurate[v]. This belief proved to be imprecise, especially when analyzing structured data with meaningful features [91]. This also depends on the algorithms being compared. For example, according to [91] it would not be fair to compare the 1984 decision tree algorithm to a more recent DL one and conclude that interpretable by design models are not as accurate. Indeed, the recently developed interpretable by design 'this looks like that' algorithm, derived from a CNN, proved to be as accurate as the non-explainable CNN [92]. Figure 5 highlights the two categories of X-AI, their corresponding representative algorithms, and the explainable outcomes associated with their implementation.

Regardless of whether a post-hoc or an interpretable by design algorithm is used, model explanations can occur on a global or local level. While the former describes the overall extracted relationships based on the entire model behavior, the latter reveals the rationale behind a specific prediction [93].

Finally, it is worth noting that, just as different X-AI techniques exist, there exists a range of approaches to explainability since different contexts give rise to different explainability needs [94]. For example, when training and evaluating an AI model, plant scientists might want to understand which data features are being used for prediction and how they are correlated together, while data scientists might require technical details about how the model functions to help in its testing, debugging, bias identification and mitigation, hyperparameter tuning, and evaluation; IS scientists can leverage details about the model training process to help optimize the architecture of the algorithm using suitable design approaches and methods (Figure 4B). At the model deployment step, regulators might require assurance about how data is being processed to assess its risk level by inspecting its reliability, as well as the impact of its predictions on its users to ultimately decide whether or not it requires authorization and regulation. Similarly, farmers and breeders might require explanations to understand why and how the model came to a prediction and to ensure its trustworthiness.

Presently, the hope for human-comprehensible explanations for black-box algorithms to increase technical confidence, generate trust, and make better informed choices remains an open challenge. In light of this challenge, we strongly recommend that a single prediction might therefore need to be explained in various ways, reflecting the requirements of all stakeholders.

How to devise X-AI-driven analytics for phenomics questions

Interpretable by design models

X-AI bears great potential for the analysis and interpretation of phenomic data. In what follows, we provide an example of an X-AI workflow design and describe for the first time, the steps needed to foster practical applicability of interpretable by design algorithms in phenomics image analysis (Figure 6). We have also accompanied this review by an interactive online tutorial that acts as an educational resource, intended for readers with little to no knowledge of X-AI algorithms; it also serves as a good starting point for self-learning and raises an early awareness that computational phenomics need not be intimidating. In addition, we have created a set of self-test quizzes and hands-on practice exercises to provide users with opportunities to augment their learning by practically applying the concepts explained in order to assess their acquired knowledge. The code and computational **notebooks** are open source and freely accessible through our GitHub repository[vi]. Collectively, this will accelerate the rate of discovery and move toward open science and AI ethics in digital phenomics.

In our tutorial, we train 'this looks like that' algorithm to classify diseases using the **crowdsourced** cassava disease classification dataset. This choice is motivated by the importance of cassava, being a key food security crop grown by smallholder farmers in Africa, Asia, and South America. However, diseases that plague the crop are a major cause of poor yield [95]. Existing methods to identify diseases require governmental agricultural experts to visually inspect and diagnose the plants [96]. This labor-intensive process makes it difficult to monitor and treat disease progression. With the help of X-AI, we can identify and classify cassava diseases and monitor their progression rapidly enough to address these current limitations in disease surveillance. Our model, initialized with **transfer learning** (TL), was trained to predict five classes, and provide corresponding prototypical explanations by marking activated patches with bounding boxes and generating heatmaps to show which parts of the image are similar to the prototypes. The resulting **confusion matrix** illustrates the percentage of correctly classified images in each class. The overall accuracy was 88.7% after cycling through the training set 240 times (Figure 6).

A more detailed description of all steps of the analysis, including the computational notebook and code to train, validate, and test/replicate our models, is provided on the tutorial website. This description covers, as relevant, data preprocessing, image classes and format, architecture of the model, model training and evaluation, prototypical explanations, and the **computer cluster** used for training.

Dealing with small datasets

As some phenotyping experiments generate small amounts of data, X-AI models get fewer training examples to learn from. But how can models learn well from small datasets? Using low complexity models that have a small number of trainable parameters can perform better than complex ones as they are less prone to overfitting and generalize better [97]. Additionally, TL can be employed to transfer knowledge acquired while learning a different but related task from a model trained on a large dataset to fit a new model using a small dataset [98]. When TL is not powerful enough due to the absence of large datasets, cumulative learning (CL) can be used to train a model over various small datasets and accumulate knowledge in the resulting network representation (i.e., model weights) [99]. Even when pretraining a model with TL or CL is not possible, the cosine loss function can substantially improve the predictive performance of the model [100]. While the loss function of the model measures the error between the input and predicted output, the cosine loss function maximizes the cosine similarity between them. One-shot or few-shot learning can also be used to train a model from one or a handful of training image data by basing predictions on a similarity metric (e.g., cosine similarity) that compares training data to new inputs [101]. Most recently, with the embedding of human knowledge into AI, it will be possible to supplement small training datasets. Representation of such knowledge can be incorporated into AI by means of changes to the input data and loss function [102], to the architecture of the algorithm [103], or to a combination thereof. Alternatively, oversampling can be another workaround that produces new sample data to augment small datasets. These new data, however, should be meaningful, sufficient, and realistic, and should contribute for better performance of predictive models [104]. Oversampling can be achieved by: (i) performing geometric transformations on existing images using primitive data manipulation techniques, including flipping, rotation, shearing, cropping, and translation [105]; (ii) generating new **synthetic data** with generative adversarial networks (GANs), which are powerful models for learning complex distributions to synthesize semantically meaningful samples from an actual training set [104]. GANs can be employed for image-to-image translation, fusion image generation, label-to-image mapping, and text-to-image translation [104]; (iii) simulating real-world scenarios, by making use of virtual reality [106] or other extended reality technologies, including augmented and mixed reality, to create immersive 3D virtual environments, in which cameras can automatically collect photorealistic synthetic images; and (iv) pairing existing images using methods such as cut-and-paste [107] or CutMix [108], which automatically ‘cut’ objects of interest from training images and ‘paste’ them on random backgrounds or on other training images, respectively.

How to ensure that human-centric X-AI benefits all: team science, open science, open education, and embedded ethics

AI in phenomics can potentially impact many aspects of plant science, from basic research discovery to translational research. It is critical that these advances in technology broadly benefit society as a whole.

So, how do we effectively ensure that human-centric X-AI benefits and does not harm individuals and communities? This can be done in several ways.

First, we suggest that pivoting toward multidisciplinary team science is necessary to tackle the most pressing scientific, societal, and ethical problems of plant digital phenomics. Over the last decade, funding agencies across the US and Europe dedicated resources to facilitating team science. This work is evidenced by interdisciplinary and multidisciplinary team requirements in funding announcements and programs. For example, addressing the problem of bias in phenomics AI requires the integrated knowledge of socially and intellectually diverse researchers who specialize in plant science, plant phenomics, plant pathology, data science, computer science, IS, social science, and bioethics, just to name a few.

Second, we emphasize the crucial importance of an open science system that aspires to open access not only to research outputs, but the whole research process, and posit that all phenomics and data centers should participate in these practices. Promotion of open science and team science are synergistic goals, both of which are essential for improving our knowledge and scientific rigor.

Third, we call for mobilizing open educational resources relevant to AI in phenomics that advocate digitized materials offered freely and openly for educators, students, and interested learners worldwide, including developing countries to use and reuse for teaching, learning, training, and research. Open education holds great promise to create knowledge and put it to use, promote content quality through sharing of materials for feedback and continuous improvement, and achieve competencies.

Fourth, we propose the development of socially and ethically responsible AI in phenomics by reforming curricula and embedding bioethicists into the technology development team. Ethical concerns around AI regarding handling of data, data bias, transparency, explainability, and responsibility, have prompted us to consider how AI technology can be designed, implemented, deployed, and monitored post deployment in an ethical manner. Embedding bioethicists into the AI development team can ensure that developers be practically assisted in anticipating, identifying, and addressing ethical issues through critical ethical reasoning and bioethical decision-making. Universities across the US and Europe have recently joined the effort to develop socially responsible AI by reforming curricula. For example, Harvard University initiated an ‘Embedded EthiCS’ curriculum that integrates ethical issues into the core computer science curriculum. We advocate that universities around the world implement similar approaches to empower students and early-career scientists to think ethically as they develop algorithms and build AI systems, in their studies, in their new business ventures, and as they pursue technical work in their careers. These free and open courses should be taught by interdisciplinary teams of computer scientists, social scientists, and bioethicists.

We encourage the scientific community to embrace a growth mindset regarding team science, open science, open education, and embedded ethics, which altogether can be harnessed to create extraordinary phenomic resources that benefit all. The rewards to these efforts come from investments of energy, time, and action.

Concluding remarks and future perspectives

We are experiencing an unprecedented time where the availability of vast amounts of phenomic data, combined with advances in AI, is providing the opportunity to turbocharge the data to insight journey. This opportunity is an incentive to not only design and implement effective and reliable data management strategies but also to improve visibility, accessibility, and usability of publicly available datasets that can support research and innovation in plant digital phenomics.

Although AI has demonstrated impressive potential in phenomics, risks due to bias and lack of transparency of models should be considered. Reducing these risks entails multidisciplinary science and technology teams working together. The involvement of plant scientists, data scientists, and IS scientists during the complete lifecycle of AI analysis is integral to ensure explainability and to identify bias in the predictive models. Interpretable by design models can potentially be leveraged to mitigate bias and provide transparency into the decision-making process.

In the past, AI research focused on a one-way interaction, from AI to humans; today, human-centric X-AI aims to enable bidirectional interaction so that human intelligence and AI are brought together to collectively achieve superior results and continuously improve by learning from each other. Human-centric X-AI will have an extraordinary impact on phenomics in the near future, and we should do all we can to ensure that it is designed, implemented, deployed, and regulated in a way that maximizes benefits for breeders, farmers, and consumers. In this regard, the academic and AI communities should ensure that computational phenomics, in addition to social and ethical analysis, are integrated into plant science curriculum as a step toward this goal (see Outstanding Questions).

Acknowledgements

The authors are grateful to the editor and three anonymous reviewers for their constructive and insightful comments which greatly helped improve the manuscript. Partial support for this work was provided by the EU FP7 project WATBIO, grant no. 311929; the EU H2020 project EMPHASIS-PREP and its Italian node, PHEN-ITALY, grant no. 739514; the Italian Ministry of University and Research Brain Gain Professorship to A.L.H; the Center for Bioenergy Innovation, a U.S. Department of Energy (DOE) Bioenergy Research Center, the Plant Microbe Interface SFA, and the integrated Pennycress Resilience Project, all supported by the Biological and Environmental Research in the DOE Office of Science; the Oak Ridge Leadership Computing Facility, a DOE Office of Science User Facility supported under Contract DE-AC05-00OR22725; and the DOE, Laboratory Directed Research and Development funding ORNL AI Initiative ProjectID 10875 at the Oak Ridge National Laboratory.

Resources

[i]https://www.top500.org/lists/top500/2022/06/

[ii]https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf

[iii]https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206

[iv]https://eur-lex.europa.eu/legal-content/en/TXT/?qid=1593079180383&uri=CELEX%3A52020DC0064

[v]https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf

[vi]https://github.com/HarfoucheLab/a-primer-on-AI-in-plant-digital-phenomics

References

1. Yang, W. *et al.* (2020) Crop phenomics and high-throughput phenotyping: past decades, current challenges, and future perspectives. *Mol. Plant* 13, 187–214
2. Houle, D. *et al.* (2010) Phenomics: the next challenge. *Nat. Rev. Genet.* 11, 855–866
3. Tardieu, F. *et al.* (2017) Plant phenomics, from sensors to knowledge. *Curr. Biol.* 27, 770–783
4. Dhondt, S. *et al.* (2013) Cell to whole-plant phenotyping: the best is yet to come. *Trends Plant Sci.* 18, 428–439

5. Fiorani, F. and Schurr, U. (2013) Future scenarios for plant phenotyping. *Annu. Rev. Plant Biol.* 64, 267–291

6. Furbank, R.T. and Tester, M. (2011) Phenomics – technologies to relieve the phenotyping bottleneck. *Trends Plant Sci.* 16, 635–644

7. Song, P. *et al.* (2021) High-throughput phenotyping: breaking through the bottleneck in future crop breeding. *Crop J.* 9, 633–645

8. Huang, Y. *et al.* (2020) Phenotypic techniques and applications in fruit trees: a review. *Plant Methods* 16, 107

9. Feng, L. *et al.* (2021) A comprehensive review on recent applications of unmanned aerial vehicle remote sensing with various sensors for high-throughput plant phenotyping. *Comput. Electron. Agric.* 182, 106033

10. Coppens, F. *et al.* (2017) Unlocking the potential of plant phenotyping data through integration and data-driven approaches. *Curr. Opin. Syst. Biol.* 4, 58–63

11. Nakhle, F. and Harfouche, A.L. (2021) Ready, steady, go AI: a practical tutorial on fundamentals of artificial intelligence and its applications in phenomics image analysis. *Patterns* 2, 100323

12. Nevavuori, P. *et al.* (2019) Crop yield prediction with deep convolutional neural networks. *Comput. Electron. Agric.* 163, 104859

13. Gopal, G. *et al.* (2019) Identification of plant leaf diseases using a nine-layer deep convolutional neural network. *Comput. Electr. Eng.* 76, 323–338

14. Smith, A.G. *et al.* (2020) Segmentation of roots in soil with U-Net. *Plant Methods*

16, 13

15. Yasrab, R. *et al.* (2019) RootNav 2.0: deep learning for automatic navigation of complex plant root architectures. *Gigascience* 8, giz123

16. Gaggion, N. *et al.* (2021) ChronoRoot: high-throughput phenotyping by deep segmentation networks reveals novel temporal parameters of plant root system architecture. *Gigascience* 10, giab052

17. Ubbens, J.R. and Stavness, I. (2017) Deep plant phenomics: a deep learning platform for complex plant phenotyping tasks. *Front. Plant Sci.* 8, 1190

18. Ghosal, S. *et al.* (2018) An explainable deep machine vision framework for plant stress phenotyping. *Proc. Natl. Acad. Sci.* 115, 4613–4618

19. Qiu, Z. *et al.* (2021) Estimation of nitrogen nutrition index in rice from UAV RGB images coupled with machine learning algorithms. *Comput. Electron. Agric.* 189, 106421

20. Shi, P. *et al.* (2021) Rice nitrogen nutrition estimation with RGB images and machine learning methods. *Comput. Electron. Agric.* 180, 105860

21. Resente, G. *et al.* (2021) Mask, train, repeat! Artificial intelligence for quantitative wood anatomy. *Front. Plant Sci.* 12, 767400

22. Nagasubramanian, K. *et al.* (2019) Plant disease identification using explainable 3D deep learning on hyperspectral images. *Plant Methods* 15, 98

23. Shu, M. *et al.* (2021) The application of UAV-based hyperspectral imaging to estimate crop traits in maize inbred lines. *Plant Phenomics* 2021, 1–14

24. Bodner, G. *et al.* (2018) Hyperspectral imaging: a novel approach for plant root phenotyping. *Plant Methods* 14, 84

25. Burnett, A.C. *et al.* (2021) Detection of the metabolic response to drought stress using hyperspectral reflectance. *J. Exp. Bot.* 72, 6474–6489

26. Tauro, F. *et al.* (2022) Latent heat flux variability and response to drought stress of black poplar: A multi-platform multi-sensor remote and proximal sensing approach to relieve the data scarcity bottleneck. *Remote Sens. Environ.* 268, 112771

27. Gené-Mola, J. *et al.* (2020) Fruit detection, yield prediction and canopy geometric characterization using LiDAR with forced air flow. *Comput. Electron. Agric.* 168, 105121

28. Raza, S.-A. *et al.* (2014) Automatic detection of regions in spinach canopies responding to soil moisture deficit using combined visible and thermal imagery. *PLoS One* 9, e97612

29. Gutiérrez, S. *et al.* (2018) Vineyard water status assessment using on-the-go thermal imaging and machine learning. *PLoS One* 13, e0192037

30. Li, J. *et al.* (2021) DeepLearnMOR: a deep-learning framework for fluorescence image-based classification of organelle morphology. *Plant Physiol.* 186, 1786–1799

31. Soltaninejad, M. *et al.* (2020) Three dimensional root CT segmentation using multi-resolution encoder-decoder networks. *IEEE Trans. Image Process.* 29, 6667–6679

32. Maimaitijiang, M. *et al.* (2020) Soybean yield prediction from UAV using multimodal data fusion and deep learning. *Remote Sens. Environ.* 237, 111599

33. Maimaitijiang, M. *et al.* (2017) Unmanned aerial system (UAS)-based phenotyping of soybean using multi-sensor data fusion and extreme learning machine. *ISPRS J. Photogramm. Remote Sens.* 134, 43–58

34. Streich, J. *et al.* (2020) Can exascale computing and explainable artificial intelligence applied to plant biology deliver on the United Nations sustainable development goals? *Curr. Opin. Biotechnol.* 61, 217–225

35. Harfouche, A.L. *et al.* (2019) Accelerating climate resilient plant breeding by applying next-generation artificial intelligence. *Trends Biotechnol.* 37, 1217–1235

36. Chhetri, H.B. *et al.* (2020) Genome-wide association study of wood anatomical and morphological traits in *Populus trichocarpa*. *Front. Plant Sci.* 11, 545748

37. Chhetri, H.B. *et al.* (2019) Multitrait genome-wide association analysis of *Populus trichocarpa* identifies key polymorphisms controlling morphological and physiological traits. *New Phytol.* 223, 293–309

38. Poorter, H. *et al.* (2016) Pampered inside, pestered outside? differences and similarities between plants growing in controlled conditions and in the field. *New Phytol.* 212, 838–855

39. Wilkinson, M.D. *et al.* (2016) The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* 3, 160018

40. Williamson, H.F. *et al.* (2021) Data management challenges for artificial

intelligence in plant and agricultural research. *F1000Research* 10, 324

41. Papoutsoglou, E.A. *et al.* (2020) Enabling reusability of plant phenomic datasets with MIAPPE 1.1. *New Phytol.* 227, 260–273

42. Ćwiek-Kupczyńska, H. *et al.* (2016) Measures for interoperability of phenotypic data: minimum information requirements and formatting. *Plant Methods* 12, 44

43. Wolf, M. *et al.* (2021) Reusability first: toward FAIR workflows. In *2021 IEEE International Conference on Cluster Computing (CLUSTER)*, pp. 444–455

44. McMahan, B. *et al.* (2017) Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 54, pp. 1273–1282

45. Chen, M. *et al.* (2021) Communication-efficient federated learning. *Proc. Natl. Acad. Sci.* 118, e2024789118

46. Kaissis, G.A. *et al.* (2020) Secure, privacy-preserving and federated machine learning in medical imaging. *Nat. Mach. Intell.* 2, 305–311

47. Kaissis, G. *et al.* (2021) End-to-end privacy preserving deep learning on multi-institutional medical imaging. *Nat. Mach. Intell.* 3, 473–484

48. Dayan, I. *et al.* (2021) Federated learning for predicting clinical outcomes in patients with COVID-19. *Nat. Med.* 27, 1735–1743

49. Brink, S.C. (2021) 25 years of trends in plant science: we should all be plant worshippers. *Trends Plant Sci.* 26, 527–529

50. Blei, D.M. and Smyth, P. (2017) Science and data science. *Proc. Natl. Acad. Sci.*

114, 8689–8692

51. Xu, L. *et al.* (2015) Plant photosynthesis phenomics data quality control. *Bioinformatics* 31, 1796–1804

52. Miceli, M. *et al.* (2020) Between subjectivity and imposition: power dynamics in data annotation for computer vision. *Proc. ACM Human-Computer Interact.* 4, 1–25

53. Akter, S. *et al.* (2021) Algorithmic bias in data-driven innovation in the age of AI. *Int. J. Inf. Manage.* 60, 102387

54. Boyd, D. and Crawford, K. (2012) Critical questions for big data. *Information, Commun. Soc.* 15, 662–679

55. Mehta, P. *et al.* (2019) A high-bias, low-variance introduction to Machine Learning for physicists. *Phys. Rep.* 810, 1–124

56. Greener, J.G. *et al.* (2022) A guide to machine learning for biologists. *Nat. Rev. Mol. Cell Biol.* 23, 40–55

57. Schwartz, R. *et al.* (2021) A proposal for identifying and managing bias in artificial intelligence. *NIST Spec. Publ.* Doi: 10.6028/NIST.SP.1270-draft

58. Großkinsky, D.K. *et al.* (2015) Plant phenomics and the need for physiological phenotyping across scales to narrow the genotype-to-phenotype knowledge gap. *J. Exp. Bot.* 66, 5429–5440

59. Matthews, D. (2018) Supercharge your data wrangling with a graphics card. *Nature* 562, 151–152

60. Mann, A. (2020) Core concept: nascent exascale supercomputers offer promise, present challenges. *Proc. Natl. Acad. Sci.* 117, 22623–22625

61. Skibba, R. (2021) Japan's fugaku supercomputer crushes competition, but likely not for long. *Engineering* 7, 6–7

62. Marx, V. (2021) Biology begins to tangle with quantum computing. *Nat. Methods* 18, 715–719

63. Dart, E. *et al.* (2014) The science DMZ: a network design pattern for data-intensive science. *Sci. Program.* 22, 701405

64. Phillips, S.J. *et al.* (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecol. Appl.* 19, 181–197

65. Castro, D.C. *et al.* (2020) Causality matters in medical imaging. *Nat. Commun.* 11, 3673

66. Buda, M. *et al.* (2018) A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks* 106, 249–259

67. Lever, J. *et al.* (2016) Regularization. *Nat. Methods* 13, 803–804

68. Jin, L. *et al.* (2020) Deep-learning-assisted detection and segmentation of rib fractures from CT scans: Development and validation of FracNet. *EBioMedicine* 62, 103106

69. Sreeram, M. and Nof, S.Y. (2021) Human-in-the-loop: role in cyber physical agricultural systems. *Int. J. Comput. Commun. Control* 16, 4166

70. Budd, S. *et al.* (2021) A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Med. Image Anal.* 71, 102062

71. Patel, B.N. *et al.* (2019) Human–machine partnership with artificial intelligence for chest radiograph diagnosis. *npj Digit. Med.* 2, 111

72. Basu, S. *et al.* (2018) Iterative random forests to discover predictive and stable high-order interactions. *Proc. Natl. Acad. Sci.* 115, 1943–1948

73. Cliff, A. *et al.* (2019) A high-performance computing implementation of iterative random forest for the creation of predictive expression networks. *Genes (Basel).* 10, 996

74. Jiang, Y. and Li, C. (2020) Convolutional neural networks for image-based high-throughput plant phenotyping: a review. *Plant Phenomics* 2020, 1–22

75. Danilevicz, M.F. *et al.* (2021) Resources for image-based high-throughput phenotyping in crops and data sharing challenges. *Plant Physiol.* 187, 699–715

76. Mochida, K. *et al.* (2019) Computer vision-based phenotyping for improvement of plant productivity: a machine learning perspective. *Gigascience* 8, giy153

77. Vilone, G. and Longo, L. (2020) Explainable artificial intelligence: a systematic review. *arXiv* 2006.00093

78. Shortliffe, E.H. *et al.* (1973) An artificial intelligence program to advise physicians regarding antimicrobial therapy. *Comput. Biomed. Res.* 6, 544–560

79. Doyle, J. (1979) A truth maintenance system. *Artif. Intell.* 12, 231–272

80. Tickle, A.B. *et al.* (1998) The truth will come to light: directions and challenges in

extracting the knowledge embedded within trained artificial neural networks. *IEEE Trans. Neural Networks* 9, 1057–1068

81. Lipton, Z.C. (2018) The mythos of model interpretability. *Queue* 16, 31–57

82. Molnar, C. (2019) *Interpretable machine learning. A guide for making black box models explainable*, (1st edition), leanpub.com.

83. Ribeiro, M.T. *et al.* (2016) “Why should I trust you?”: explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144

84. Lundberg, S.M. and Lee, S.I. (2017) A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, 30, pp. 4768–4777

85. Ghosal, S. *et al.* (2018) An explainable deep machine vision framework for plant stress phenotyping. *Proc. Natl. Acad. Sci.* 115, 4613–4618

86. Toda, Y. and Okura, F. (2019) How convolutional neural networks diagnose plant disease. *Plant Phenomics* 2019, 1–14

87. Nagasubramanian, K. *et al.* (2019) Plant disease identification using explainable 3D deep learning on hyperspectral images. *Plant Methods* 15, 98

88. Nagasubramanian, K. *et al.* (2020) Usefulness of interpretability methods to explain deep learning based plant stress phenotyping. *arXiv* 2007.05729

89. Mostafa, S. *et al.* (2021) Visualizing feature maps for model selection in convolutional neural networks. In *Proceedings of the IEEE/CVF International*

Conference on Computer Vision (ICCV) Workshops, pp. 1362–1371

90. Dobrescu, A. *et al.* (2019) Understanding deep neural networks for regression in leaf counting. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2600–2608

91. Rudin, C. (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* 1, 206–215

92. Chen, C. *et al.* (2019) This looks like that: deep learning for interpretable image recognition. In *Advances in Neural Information Processing Systems*, 32, pp. 8930–8941

93. Barredo Arrieta, A. *et al.* (2020) Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* 58, 82–115

94. Preece, A. *et al.* (2018) Stakeholders in explainable AI. *arXiv* 1810.00184

95. Otim-Napr, G.W. *et al.* (2001) Changes in the incidence and severity of Cassava mosaic virus disease, varietal diversity and cassava production in Uganda. *Ann. Appl. Biol.* 138, 313–327

96. Mwebaze, E. *et al.* (2019) iCassava 2019 fine-grained visual categorization challenge. *arXiv* 1908.02900

97. Brigato, L. and Iocchi, L. (2021) A close look at deep learning with small data. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 2490–

2497

98. Gupta, V. *et al.* (2021) Cross-property deep transfer learning framework for enhanced predictive analytics on small materials data. *Nat. Commun.* 12, 6595

99. Seddiki, K. *et al.* (2020) Cumulative learning enables convolutional neural network representations for small mass spectrometry data classification. *Nat. Commun.* 11, 5595

100. Barz, B. and Denzler, J. (2020) Deep learning on small datasets without pre-training using cosine loss. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1360–1369

101. Wang, Y. *et al.* (2021) Generalizing from a few examples: a survey on few-shot learning. *ACM Comput. Surv.* 53, 1–34

102. Childs, C.M. and Washburn, N.R. (2019) Embedding domain knowledge for machine learning of complex material systems. *MRS Commun.* 9, 806–820

103. Hasebe, T. (2021) Knowledge-embedded message-passing neural networks: improving molecular property prediction with human knowledge. *ACS Omega* 6, 27955–27967

104. Shamsolmoali, P. *et al.* (2021) Image synthesis with adversarial networks: a comprehensive survey and case studies. *Inf. Fusion* 72, 126–146

105. Khalifa, N.E. *et al.* (2022) A comprehensive survey of recent trends in deep learning for digital images augmentation. *Artif. Intell. Rev.* 55, 2351–2377

106. Martinez-Gonzalez, P. *et al.* (2020) UnrealROX: an extremely photorealistic virtual

reality environment for robotics simulations and synthetic data generation. *Virtual Real.* 24, 271–288

107. Dwibedi, D. *et al.* (2017) Cut, paste and learn: surprisingly easy synthesis for instance detection. In *2017 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1310–1319

108. Yun, S. *et al.* (2019) CutMix: regularization strategy to train strong classifiers with localizable features. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6022–6031

109. Pound, M.P. *et al.* (2017) Deep machine learning provides state-of-the-art performance in image-based plant phenotyping. *Gigascience* 6, gix083

110. Pound, M.P. *et al.* (2017) Deep learning for multi-task plant phenotyping. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 2055–2063

111. David, E. *et al.* (2020) Global wheat head detection (GWHD) dataset: a large and diverse dataset of high-resolution RGB-labelled images to develop and benchmark wheat head detection methods. *Plant Phenomics* 2020, 1–12

112. David, E. *et al.* (2021) Global wheat head detection 2021: an improved dataset for benchmarking wheat head detection methods. *Plant Phenomics* 2021, 1–9

113. Wang, X. *et al.* (2019) High-throughput phenotyping with deep learning gives insight into the genetic architecture of flowering time in wheat. *Gigascience* 8, 1–11

114. Quiñones, R. *et al.* (2021) Multi-feature data repository development and analytics for image cosegmentation in high-throughput plant phenotyping. *PLoS One* 16, e0257001

115. Taghavi Namin, S. *et al.* (2018) Deep phenotyping: Deep learning for temporal phenotype/genotype classification. *Plant Methods* 14, 66

116. Minervini, M. *et al.* (2016) Finely-grained annotated datasets for image-based plant phenotyping. *Pattern Recognit. Lett.* 81, 80–89

117. Cruz, J.A. *et al.* (2016) Multi-modality imagery database for plant phenotyping. *Mach. Vis. Appl.* 27, 735–749

118. Dobos, O. *et al.* (2019) A deep learning-based approach for high-throughput hypocotyl phenotyping. *Plant Physiol.* 181, 1415–1424

119. Khanna, R. *et al.* (2019) A spatio temporal spectral framework for plant stress phenotyping. *Plant Methods* 15, 13

120. Sa, I. *et al.* (2018) WeedMap: a large-scale semantic weed mapping framework using aerial multispectral imaging and deep neural network for precision farming. *Remote Sens.* 10, 1423

121. Chebrolu, N. *et al.* (2017) Agricultural robot dataset for plant classification, localization and mapping on sugar beet fields. *Int. J. Rob. Res.* 36, 1045–1052

122. Mignoni, M.E. *et al.* (2022) Soybean images dataset for caterpillar and *Diabrotica speciosa* pest detection and classification. *Data Br.* 40, 107756

123. dos Santos Ferreira, A. *et al.* (2017) Weed detection in soybean crops using

ConvNets. *Comput. Electron. Agric.* 143, 314–324

124. Bosilj, P. *et al.* (2020) Transfer learning between crop types for semantic segmentation of crops versus weeds in precision agriculture. *J. F. Robot.* 37, 7–19

125. Haug, S. and Ostermann, J. (2015) A crop/weed field image dataset for the evaluation of computer vision based precision agriculture tasks. In *Computer Vision - ECCV 2014 Workshops*, pp. 105–116

126. Nakatumba-Nabende, J. *et al.* (2020) A dataset of necrotized cassava root cross-section images. *Data Br.* 32, 106170

127. Thapa, R. *et al.* (2020) The plant pathology challenge 2020 data set to classify foliar disease of apples. *Appl. Plant Sci.* 8, e11390

128. Gené-Mola, J. *et al.* (2020) LFuji-air dataset: annotated 3D LiDAR point clouds of Fuji apple trees for fruit detection scanned under different forced air flow conditions. *Data Br.* 29, 105248

129. Hani, N. *et al.* (2020) MinneApple: a benchmark dataset for apple detection and segmentation. *IEEE Robot. Autom. Lett.* 5, 852–858

130. Dias, P.A. *et al.* (2018) Multispecies fruit flower detection using a refined semantic segmentation network. *IEEE Robot. Autom. Lett.* 3, 3003–3010

131. Fenu, G. and Malloci, F.M. (2021) DiaMOS plant: a dataset for diagnosis and monitoring plant disease. *Agronomy* 11, 2107

132. Bargoti, S. and Underwood, J. (2017) Deep fruit detection in orchards. In *2017*

IEEE International Conference on Robotics and Automation (ICRA), pp. 3626–3633

133. Kusrini, K. *et al.* (2020) Data augmentation for automated pest classification in Mango farms. *Comput. Electron. Agric.* 179, 105842

134. Kicherer, A. *et al.* (2015) An automated field phenotyping pipeline for application in grapevine research. *Sensors* 15, 4823–4836

135. Navarro, P.J. *et al.* (2022) A novel ground truth multispectral image dataset with weight, anthocyanins, and Brix index measures of grape berries tested for its utility in machine learning pipelines. *Gigascience* 11, giac052

136. Alessandrini, M. *et al.* (2021) A grapevine leaves dataset for early detection and classification of esca disease in vineyards through machine learning. *Data Br.* 35, 106809

137. Abdelghafour, F. *et al.* (2021) An annotated image dataset of downy mildew symptoms on Merlot grape variety. *Data Br.* 37, 107250

138. Kestur, R. *et al.* (2019) MangoNet: a deep semantic segmentation architecture for a method to detect and count mangoes in an open orchard. *Eng. Appl. Artif. Intell.* 77, 59–69

139. Chouhan, S.S. *et al.* (2019) A data repository of leaf images: practice towards plant conservation with plant pathology. In *2019 4th International Conference on Information Systems and Computer Networks (ISCON)*, pp. 700–707

140. Ali, H. *et al.* (2017) Symptom based automated detection of citrus diseases using

color histogram and textural descriptors. *Comput. Electron. Agric.* 138, 92–104

141. Sharif, M. *et al.* (2018) Detection and classification of citrus diseases in agriculture based on optimized weighted segmentation and feature selection. *Comput. Electron. Agric.* 150, 220–234

142. Chitwood, D.H. and Otoni, W.C. (2017) Morphometric analysis of *Passiflora* leaves: the relationship between landmarks of the vasculature and elliptical Fourier descriptors of the blade. *Gigascience* 6, giw008

143. Chitwood, D.H. and Otoni, W.C. (2017) Divergent leaf shapes among *Passiflora* species arise from a shared juvenile morphology. *Plant Direct* 1, e00028

144. Kaufmane, E. *et al.* (2022) QuinceSet: dataset of annotated Japanese quince images for object detection. *Data Br.* 42, 108332

145. Altaheri, H. *et al.* (2019) Date fruit dataset for intelligent harvesting. *Data Br.* 26, 104514

146. Altaheri, H. *et al.* (2019) Date fruit classification for robotic harvesting in a natural environment using deep learning. *IEEE Access* 7, 117115–117133

147. Wiesner-Hanks, T. *et al.* (2018) Image set for deep learning: field images of maize annotated with disease symptoms. *BMC Res. Notes* 11, 440

148. Lac, L. *et al.* (2022) An annotated image dataset of vegetable crops at an early stage of growth for proximal sensing applications. *Data Br.* 42, 108035

149. Schunck, D. *et al.* (2021) Pheno4D: a spatio-temporal dataset of maize and tomato plant point clouds for phenotyping and advanced plant analysis. *PLoS*

One 16, e025634

150. Jepkoech, J. *et al.* (2021) Arabica coffee leaf images dataset for coffee leaf disease detection and classification. *Data Br.* 36, 107142

151. Manso, G.L. *et al.* (2019) A smartphone application to detection and classification of coffee leaf miner and coffee leaf rust. *arXiv* 1904.00742

152. Parraga-Alava, J. *et al.* (2019) RoCoLe: a robusta coffee leaf images dataset for evaluation of machine learning based methods in plant diseases recognition. *Data Br.* 25, 104414

153. Uchiyama, H. *et al.* (2017) An easy-to-setup 3D phenotyping platform for KOMATSUNA dataset. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 2038–2045

154. Seidel, D. *et al.* (2021) Predicting tree species from 3D laser scanning point clouds using deep learning. *Front. Plant Sci.* 12, 1–12

155. Espejo-Garcia, B. *et al.* (2020) Towards weeds identification assistance through transfer learning. *Comput. Electron. Agric.* 171, 105306

156. Olsen, A. *et al.* (2019) DeepWeeds: a multiclass weed species image dataset for deep learning. *Sci. Rep.* 9, 2058

Figure legends

Figure 1. Phenomics platforms and sensors for high-throughput plant phenotyping in controlled environments and field conditions: collecting relevant data from a wide range of sources. A network of comprehensive automated weather stations collects hourly

weather and soil data, including, among others, rainfall, air temperature, solar radiation, relative humidity, and soil moisture and temperature. Varying phenotyping scales allow for precise and consistent monitoring of individual plants, plots, and fields. Ground-based and aerial platforms can mount a variety of cameras and sensors for non-invasive, high-throughput (HTP) phenotyping: visible light camera for RGB imaging; LiDAR sensor and 3D laser scanners for 3D imaging; multispectral cameras and hyperspectral sensors for spectral imaging; TIR cameras for thermal imaging; and chlorophyll fluorescence sensor for chlorophyll fluorescence imaging. Automated and environmentally controlled platforms, growth chambers, and multifunction printers can be used for HTP in controlled environments. Root phenotyping in the field can be invasive (e.g., shovelomics and its automation with root excavating robots); minimally invasive (e.g., minirhizotrons); or non-invasive (e.g., ERT, electrical capacitance, GPR mapping, and electromagnetic inductance mapping). Field deployable linear X-ray CT cart, and handheld X-ray fluorescence elemental mapping are being explored for non-invasive field root phenotyping. Multispectral, hyperspectral, RGB, and EIT imaging can be used to phenotype roots in soil-filled rhizotrons (rhizoboxes) in controlled environments. Similarly, NMR, X-ray CT, and PET imaging can be used to phenotype roots in soil-filled pots. RhizoTubes, which are cylindrical rhizotrons, allow full visualization of the root system of a single or up to six plants simultaneously. The RhizoCab is designed to take images of the entire root systems of plants growing in RhizoTubes. These platforms and sensing technologies are generating a massive amount of data, which creates a need for proper data management and processing – the first step of the data life cycle in digital phenomics (Figure 2). Abbreviations: EIT, electrical impedance tomography; ERT, electrical resistance tomography; GPR, ground penetrating radar; LiDAR, light detection and ranging; NMR, nuclear magnetic resonance; PET, positron emission tomography; RGB, red–green–blue; TIR, thermal infrared; UAVs, unmanned aerial vehicles; X-ray CT, X-ray computed tomography.

Figure 2. Data architecture blueprint to drive human-centric explainable artificial intelligence (X-AI) innovation. Phenomics data can be structured, semistructured, or unstructured. Structured (e.g., spreadsheet files) data (blue line) are typically 'at-rest', transformed into rows and columns, and loaded into relational databases in data warehouses using a process known as ETL. Semistructured (e.g., extensible markup language files) and unstructured (e.g., flat files) data (red line) are streamed 'in-motion', loaded into non-relational databases, and stored in data lakes in their raw form; their transformation occurs on-demand using a process known as ELT. As ELT loads data immediately, it prevents any slowdown that often occurs at the transformation step, and thus, enables near real-time analytics for fast and practical decision-making. Whether ETL or ELT is used, data warehouses and data lakes store data as matrices, cubes, polytopes, or distributed in memory. A well-designed data architecture results in higher-quality phenomic datasets that allow plant scientists to ask biological questions and to devise data-driven analytics, searching for answers. Abbreviations: ELT, extract, load, transform; ETL, extract, transform, load.

Figure 3. Technology infrastructure to support human-centric explainable artificial intelligence (X-AI). The technology infrastructure consists of the hardware, network flow, software frameworks, and programming languages that enable data transmission, transformation, storage, access, and analysis. (A) Computing hardware supporting data

analysis. Pre-exascale supercomputers (e.g., University of Waterloo's Graham, and Lawrence Livermore National Laboratory's Sierra) reach a performance of a million billion FLOPS. With a similar hardware architecture but an increased number of CPUs and GPUs, exascale supercomputers (e.g., Oak Ridge National Laboratory's Summit and Frontier) reach a billion billion FLOPS and can deliver higher performance in pattern searching in phenomic big data, and thus, speeding up crop design. Quantum computers (e.g., International Business Machines' System One and Quantinuum's H1-2) represent a new paradigm in computation that leverages the fundamental principles of quantum mechanics to perform calculations. They employ **quantum bits** (qubits) that can be entangled, giving them the ability to manipulate vast amounts of data with few operations, and thus, the capacity to solve problems polynomially faster than classical computers (i.e., pre-exascale and exascale supercomputers) to ultimately design faster, better crops. Researchers can simulate quantum circuits on classical computers using free and open-source software development kits such as Cirq or Qiskit, and the cuQuantum software library to leverage the power of GPUs and parallel computing to perform faster calculations. Examples of classical and quantum computers are compared based on their peak performance that is the theoretical highest processing power they can reach. For classical computers, the LINPACK benchmark tests the performance in double precision (64-bit) compute capabilities while HPL-AI scores performance based on mixed precision (16- and 32-bit). As quantum computers use **QPUs** to manipulate the quantum states of qubits to perform computations, their performance is measured using **QV**. (B) Network flow to enable high-throughput access to and sharing of phenomic datasets. Requests coming from the wide area network are forwarded through a router to one of two paths: (i) the data query and browse path (red line) where requests to browse or search phenomic datasets are filtered through a firewall and processed by the hosting server; and (ii) the data transfer path (green line) where requests to download or upload phenomic datasets are inspected in the DMZ for access control, and are forwarded to the transfer nodes (typically Linux servers) to reach the filesystem where data can be transformed before or after storage using ETL or ELT, respectively (see Figure 2). (C, D) Representative free and open-source software frameworks and their supported programming languages used to implement AI algorithms. Abbreviations: CPU, central processing unit; ELT, extract, load, transform; ETL, extract, transform, load; FLOPS, floating-point operations per second; GPU, graphics processing unit; HPL-AI, high performance LINPACK for accelerator introspection; LINPACK, linear equations software package; PB, petabyte; QPU, quantum processing unit; Qubit, quantum bit; QV, quantum volume.

Figure 4. Artificial intelligence (AI) architecture design to unleash the power of human-centric explainable AI (X-AI). (A) Representative AI algorithms that are used for AI tasks in digital phenomics including classification and regression (supervised learning), and clustering and dimensionality reduction (unsupervised learning). Reinforcement learning algorithms can be applied to search optimal architecture designs and improve their performance. (B) Representative AI algorithm design approaches and methods, a higher level of abstraction that help scientists in their efforts to design and implement novel AI algorithms to answer complex biological questions. The knowledge-based AI approach represents human expert knowledge as a collection of rules to form a knowledge base that is applied to solve a specific problem. It offers a consistent answer

for a repetitive problem and its decisions are explainable. It can be implemented using rule-based methods. The Data-driven AI approach discovers connections and correlations automatically in a large amount of data and learns a black box model. It can be implemented using various methods including CNN, ensemble, or statistical methods, among others. The Informed AI approach combines knowledge-based AI with data-driven AI by leveraging human knowledge with knowledge acquired from data to make faster, more accurate decisions. It can be implemented using ensemble or rule-based methods. Finally, X-AI approaches provide meaningful explanations of decisions made by X-AI models to humans through a decipherable decision-making process. They allow the monitoring of inputs and outputs with the purpose of verifying X-AI models' adherence to ethical and socio-legal values by: (i) opening the black box of data-driven or informed AI models using ensemble methods; or (ii) designing new, transparent glass box algorithms that are interpretable by design using ensemble or CNN methods. Abbreviations: CNN, convolutional neural network; DBSCAN, density-based spatial clustering of applications with noise; DNN, deep neural network; GAN, generative adversarial network; GMM, Gaussian mixture model; HMM, hidden Markov model; KNN, k-nearest neighbors; NN, neural network; PCA, principal component analysis; RNN, recurrent neural network; RF, random forest; SAE, sparse autoencoder; SARSA, state–action–reward–state–action; SSAE, stacked SAE; SVM, support vector machine.

Figure 5. Cultivating conditions for explainable artificial intelligence (X-AI) to flourish in plant digital phenomics. Data preprocessing prepares input data for X-AI algorithms: descriptive data analysis provides statistical summaries about a dataset in order to spot

anomalies; data annotation and standardization is done by labeling and adding relevant, structured information about the data such as its source and other details known as metadata; and feature engineering uses existing features to create new ones while feature selection extracts relevant features from the complete set of features in a dataset, increasing the predictive precision of learning algorithms. X-AI can be achieved by either opening the black box or designing a transparent glass box. X-AI can be interrogated to understand why a decision has been made, keeping human-in-the-loop (HITL) of such decision-making, and allowing a two-way transfer of knowledge where on the one hand, experts assist in the training of X-AI and on the other hand, explanations can be used to generate scientific hypotheses that can result in new discoveries. An X-AI that takes into account the requirements of all stakeholders interacting with it will drive successful adoption among agricultural technopreneurs, plant biologists, policymakers, and funders. This will help bridge the gap between science, policy, embedded ethics, and entrepreneurship, allowing for responsible TT, and leading to technological, regulatory, and social and ethical outcomes. Abbreviations: AI, artificial intelligence; CNN, convolutional neural network; DeconvNet, deconvolution network; DeepLift, deep learning important features; FAIR, findable, accessible, interoperable, reusable; IPP, intellectual property protection; LIME, local interpretable model-agnostic explanations; SHAP, Shapley additive explanation; TT, technology transfer.

Figure 6. Planning, training, and interpreting an explainable artificial intelligence (X-AI)-based analysis in plant digital phenomics require careful consideration at each stage of the analysis. This figure sheds light on all the elements of designing such a workflow using cassava leaf disease classification task as an example. For data preparation, (i) a

dataset shared on Kaggle by the AI lab at Makerere University was used for analysis; (ii) data cleaning was carried out to eliminate outliers and mislabeled images; (iii) the dataset was randomly split for training, validation, and testing; (iv) another shared version of the dataset with images cropped to leaf boundaries using a trained YOLO model was used to minimize noise in training images; and (v) the training dataset was augmented and balanced by oversampling, creating random transformations to image geometries. An alternative solution to oversampling is synthesizing leaf images; OpenCV can be used to segment leaves to train a deep convolutional generative adversarial network (DCGAN) to generate synthetic data. 'This looks like that' interpretable by design algorithm, implemented in Python and PyTorch was carefully chosen for the classification task; its training time was approximated and compared on different hardware, showing the advantages of GPUs over CPUs and exascale over pre-exascale supercomputers. However, increasing the number of GPUs comes at the price of increased network communication and input-output (I/O) operations to synchronize the model over cluster nodes. Such overheads can cause a delay in the training time. For example, while the algorithm is expected to complete 1000 training epochs in 31 hours using 26 Nvidia Tesla V100 GPUs, it is still expected to take an approximation of eight hours using the full power of Summit supercomputer (27,649 GPUs). 'This looks like that' algorithm uses transfer learning to import convolutional layers from pre-trained models and during training, the prototype layer extracts parts of training images (prototypes) and learns a similarity metric between them; the final class prediction is based on the weighted sum of similarities between the input and prototypes. For some prototypes, the nearest image patches come from different classes, often corresponding to a background patch, and thus should be pruned. For interpretation, the model tries to find evidence for a test image to belong to a specific class, marking activated patches by bounding boxes. While heatmaps show which part of the image is similar to a prototype, the confusion matrix illustrates the percentage of images of a true class classified into the class indicated by the predicted class column, indicating an overall accuracy of 88.7% after 240 training epochs. Abbreviations: CPU, central processing unit; DenseNet, dense convolutional network; GPU, graphics processing unit; OpenCV, open source computer vision library; ResNet, residual network; VGG, visual geometry group; YOLO, you only look once.

Table 1. Publicly available global datasets and their characteristics as valuable resources for plant digital phenomics research[a,b]

Dataset	Country of origin	Plant species[c]	Plant organ systems	No. of images	Platform	Sensors	Image annotation types[d]	Potential applications	Access type[e]	Data access details	Data PID[f]	Link to dataset	File format[g]	Refs.
Supporting data for "deep machine learning provides state-of-the-art performance in image-based plant phenotyping"	UK	*Triticum aestivum*	Root	2697	Controlled environment stationary platform	RGB	Point annotations of root tips, leaf tips, leaf bases, ear tips, ear bases	Identification and localization of root tips, leaf, and wheat ear tips	OA	Downloadable tar file	http://doi.org/10.5524/100343	http://gigadb.org/dataset/100343	JPG	[109]
			Shoot	1664										
Wheat 2017	UK	*T. aestivum*	Shoot, spikes	520	Controlled environment stationary platform	RGB	Point annotations of spikelet, base and tip of each ear	Localization and counting of wheat spikes and spikelets	OA	Register for link to download zip file	–	https://plantimages.nottingham.ac.uk/	JPG	[110]
Global wheat head detection (GWHD)	Japan, France, Canada, UK, Switzerland, China, Australia	Wheat	Shoot	1094	Manned mobile platform	RGB	Bounding boxes	Detection and localization of wheat heads	OA	Downloadable zip file	http://doi.org/10.5281/zenodo.4298502	https://zenodo.org/record/4298502	PNG	[111]
				678	Handheld visible light camera in the field									
				447	Rail-based field automated gantry									
Global wheat head detection (GWHD) 2021	Japan, France, Canada, UK, Switzerland, China, Australia, USA, Mexico, Republic of Sudan, Norway, Belgium	Wheat	Shoot	2307	Manned mobile platform	RGB	Bounding boxes	Detection and localization of wheat heads	OA	Downloadable zip file	https://doi.org/10.5281/zenodo.5092309	https://zenodo.org/record/5092309	PNG	[112]
				2684	Handheld visible light camera in the field									
				1429	Rail-based field automated gantry									

Supporting data for "high throughput phenotyping with deep learning gives insight into the genetic architecture of flowering time in wheat"	KS - USA	Wheat	Shoot	>400000	Ground-based field robot	RGB	Image-level annotations	Estimation of plant morphology and developmental stages	OA	Downloadable zip file	http://dx.doi.org/10.5524/100566	http://gigadb.org/dataset/view/id/100566	JPEG	[113]
RootNav 2.0	UK	*T. aestivum*	Root	3630	Controlled environment stationary platform	RGB	Segmentation masks, image-level annotations	Root segmentation and species classification	OA	Download each image separately	http://doi.org/10.5524/100651	http://gigadb.org/dataset/100651	JPG	[15]
		Brassica napus		120		NIR							PNG	
		Arabidopsis thaliana		277										
Cosegmentation for plant phenotyping (CosegPP)	USA	Buckwheat	Shoot	56	Controlled environment stationary platform	RGB	Segmentation masks	Plant segmentation	OA	Downloadable zip file	https://doi.org/10.5281/zenodo.5117176	https://zenodo.org/record/5117176	PNG	[114]
				56		IR								
				56		CF								
		Sunflower		104		RGB								
				112		IR								
				112		CF								
Aberystwyth leaf evaluation dataset	UK	*A. thaliana*	Shoot	56	Controlled environment stationary platform	RGB	Semantic segmentation	Plant and leaf segmentation	OA	Downloadable zip file	http://doi.org/10.5281/zenodo.168158	https://zenodo.org/record/168158	PNG	
Deep phenotyping dataset	Australia	*A. thaliana*	Shoot	2134	Controlled environment stationary platform	RGB	Image-level annotations	Genotype classification	OA	Downloadable zip file	–	https://figshare.com/s/e18a978267675059578f	JPG	[115]
Supporting data for "ChronoRoot: high-throughput phenotyping by deep segmentation networks reveals novel temporal parameters of plant root system architecture"	France	*A. thaliana*	Root	331	Controlled environment stationary platform	RGB	Segmentation masks	Root segmentation	OA	Downloadable tar file	http://dx.doi.org/10.5524/100911	http://dx.doi.org/10.5524/100911	PNG	[16]
Plant phenotyping datasets	Italy	*A. thaliana*	Shoot	6287	GARNICS controlled environment robot gardener	RGB	Segmentation masks, bounding boxes, point annotations of	Plant and leaf segmentation, leaf counting, species classification	OA	Fill in a form and get access and download a zip file	–	https://www.plant-phenotyping.org/datasets-	HDF5	[116]

Table 1. (continued)

	Germany	*Nicotiana tabacum*		165120			leaf centers					home		
Multi-modality plant imagery database (MSU-PID)	MI - USA	*A. thaliana*	Shoot	576	Controlled environment stationary platform	RGB, RGB-depth, CF, NIR[h]	Polygon and point annotations of leaves and leaf tips	Leaf segmentation, counting, alignment, and tracking	OA	Downloadable zip file	–	http://cvlab.cse.msu.edu/multi-modality-imagery-database-msu-pid.html	PNG	[117]
		Phaseolus vulgaris		175										
Plant segmentation	Hungary	*A. thaliana*	Shoot	16[i]	Computer scanner	RGB	Segmentation masks	Segmentation and length determination of hypocotyl	OA	Downloadable zip file	–	https://www.kaggle.com/tivadardanka/plant-segmentation	PNG	[118]
		Brachypodium distachyon		8[i]	Handheld visible light camera in a controlled environment									
		Sinapis alba	Shoot and root	15[i]										
Eschikon plant stress phenotyping dataset	Switzerland	*Beta vulgaris*	Shoot	496	Controlled environment stationary platform	HS (NIR)	Image-level annotations, bounding boxes	Classification of biotic and abiotic stress	OA	Downloadable zip file	–	https://projects.asl.ethz.ch/datasets/doku.php?id=2018plantstressphenotyping	PNG	[119]
				992		RGB-depth								
				496		RGB								
Remote sensing 2018 weed map dataset	Switzerland, Germany	*B. vulgaris*, weed	Shoot	18746	UAV	MS (NIR), 4 and 5 bands[j]	Semantic segmentation	Identification and segmentation of crops and weeds	OA	Downloadable zip file	–	https://projects.asl.ethz.ch/datasets/doku.php?id=weedmap:remotesensing2018weedmap	TIF	[120]
Sugar beets 2016	Germany	Sugar beet, weed	Shoot	300	Ground-based field robot	RGB, RGB-depth, MS (NIR), 4 bands[h]	Semantic segmentation	Identification and segmentation of crops and weeds	OA	Download each image separately	–	https://www.ipb.uni-bonn.de/datasets_IJRR2017/	PNG	[121]
Images of soybean leaves	Brazil	Soybean	An intact leaf, shoot	6410	Handheld visible light camera in the field, UAV	RGB	Image-level annotations	Identification and classification of biotic stress	OA	Downloadable zip file	https://doi.org/10.17632/bycbh73438.1	https://data.mendeley.com/datasets/bycbh73438/1	JPG	[122]
Data for: weed detection in soybean crops using ConvNets	Brazil	Soybean, weed	Shoot	400	UAV	RGB	Segmentation masks	Identification and segmentation of crops and weeds	OA	Downloadable zip file	https://doi.org/10.17632/3fmjm7ncc6.2	https://data.mendeley.com/datasets/3fmjm7ncc6/2	JPEG	[123]
Crop vs weed discrimination dataset	UK	Onion, weed	Shoot	20[k]	Manned mobile platform	RGB, MS (NIR), 2 bands[h]	Semantic segmentation	Identification and segmentation of crops and weeds	OA	Downloadable zip file	–	https://lcas.lincoln.ac.uk/wp/research/data-sets-software/crop-vs-weed-discrimination-dataset/	PNG	[124]
		Carrot, weed		20[k]										
Crop/weed field image dataset (CWFID)	Germany	Carrot, weed	Shoot	60	Ground-based field robot	MS (NIR), 2 bands	Segmentation masks, semantic segmentation	Identification and segmentation of crops and weeds	OA	Downloadable zip file	–	https://github.com/cwfid/dataset	PNG	[125]

Cassava leaf disease classification	Uganda	*Manihot esculenta*	Shoot	9436	Handheld visible light camera in the field	RGB	Image-level annotations	Identification and classification of biotic stress	OA	Create Kaggle account to download zip file	–	https://www.kaggle.com/c/cassava-disease/data	JPG	[96]
Cassava disease classification		Cassava		21397								https://www.kaggle.com/c/cassava-leaf-disease-classification/data		
Cassava root cross-section images	Uganda, Tanzania	Cassava	Root	10052	Handheld visible light camera in the field	RGB	Semantic segmentation	Quantification of root damage	OA	Downloadable zip file	https://doi.org/10.17632/gvp7vshvnh.3	https://data.mendeley.com/datasets/gvp7vshvnh/3	JPG	[126]
Plant pathology 2020 - FGVC7	NY - USA	Apple	An intact leaf	3651	Handheld visible light camera in the field	RGB	Image-level annotations	Identification and classification of biotic stress	OA	Create Kaggle account to download zip file	–	https://www.kaggle.com/c/plant-pathology-2020-fgvc7/data	JPG	[127]
Plant pathology 2021 – FGVC8				23000								https://www.kaggle.com/c/plant-pathology-2021-fgvc8		
LFuji-air dataset	Spain	*Malus domestica*	Shoot	88	Ground-based field robot	LiDAR	Bounding boxes	Identification and localization of fruits, estimation of yield, canopy geometric characterization	OA	Downloadable zip file	–	https://repositori.udl.cat/handle/10459.1/68782	MAT	[27,128]
MinneApple	MN - USA	Apple	Shoot	1000	Handheld visible light camera in the field	RGB	Polygon annotations	Identification, segmentation, and counting of fruits	OA	Downloadable tar file	http://doi.org/10.13020/8ecp-3r13	https://conservancy.umn.edu/handle/11299/206575	JPG	[129]
Data from: multi-species fruit flower detection using a refined semantic segmentation network	WV - USA	Apple	Shoot	18[i]	Manned mobile platform	RGB	Segmentation masks, image-level annotations	Classification of species from fruit flowers	OA	Downloadable zip file	http://doi.org/10.15482/USDA.ADC/1423466	https://data.nal.usda.gov/dataset/data-multi-species-fruit-flower-detection-using-refined-semantic-segmentation-network	JPG	[130]
				130	Handheld visible light camera in the field									
		Peach		24[i]										
		Pear		18[i]										
DiaMOS plant dataset: a dataset for diagnosis and monitoring plant disease	Italy	Pear	Shoot	3505	Handheld visible light camera in the field	RGB	Image-level annotations	Identification and classification of biotic stress	OA	Downloadable zip file	https://doi.org/10.5281/zenodo.5557313	https://zenodo.org/record/5557313#.Yv5JrXZByMo	JPG	[131]
PlantaeK: A leaf database of native plants of Jammu and Kashmir	Jammu, Kashmir	Apple	A single detached leaf	351	Handheld visible light camera in a controlled environment	RGB	Image-level annotations	Identification of biotic stress, classification of plant species	OA	Downloadable zip file	http://doi.org/10.17632/t6j2h22jpx.1	https://data.mendeley.com/datasets/t6j2h22jpx/1	JPG	
		Apricot		270										
		Cherry		212										
		Cranberry		212										

		Grapevine		171										
		Peach		331										
		Pear		228										
		Walnut		378										
Data for: identification of plant leaf diseases using a 9-layer deep convolutional neural network	FL - PA - NY - USA	Multiple crops[i]	A single detached leaf	54305	Handheld visible light camera in a controlled environment	RGB	Image-level annotations	Identification and classification of biotic stress	OA	Downloadable zip file	http://doi.org/10.17632/tywbtsjrjv.1	https://data.mendeley.com/datasets/tywbtsjrjv/1	JPG	[13]
ACFR orchard fruit dataset	Australia	Apple	Shoot	1120	Ground-based field robot	RGB	Bounding boxes, circle annotations	Identification and classification of fruits	OA	Downloadable zip file	–	http://data.acfr.usyd.edu.au/ag/treecrops/2016-multifruit/	PNG	[132]
		Almond		620										
		Mango		1964										
Dataset for pest classification in Mango farms	Indonesia	Mango	An intact leaf	510	Handheld visible light camera in the field	RGB	Image-level annotations	Classification of pest	OA	Downloadable zip file	https://doi.org/10.17632/94jf97jzc8.1	https://data.mendeley.com/datasets/94jf97jzc8/1	JPG	[133]
Berries in vineyards-color (BIVcolor)	Germany	*Vitis vinifera*	Shoot	500	Ground-based field robot	RGB	Image-level annotations	Detection of grape size and color	OA	Downloadable zip file	http://doi.org/10.5073/jki-data.2015.1	https://www.openagrar.de/receive/openagrar_mods_00021925	TIF	[134]
Supporting data for "a novel ground truth multispectral image dataset with weight, anthocyanins and Brix index measures of grape berries tested for its utility in machine learning pipelines"	Spain	*V. vinifera*	Fruit	1283	Controlled environment stationary platform	MS	Image-level annotations	Prediction of grape variety	OA	Downloadable zip file	http://dx.doi.org/10.5524/102220	http://gigadb.org/dataset/102220	TIF	[135]
ESCA-dataset	Italy	Grapevine	Shoot	1770	Handheld visible light camera in the field	RGB	Image-level annotations	Identification and classification of biotic stress	OA	Downloadable zip file	http://doi.org/10.17632/89cnxc58kj.1	https://data.mendeley.com/datasets/89cnxc58kj/1	JPG	[136]
An annotated image dataset of downy mildew symptoms on Merlot grape variety	France	Grapevine	Shoot	99	Ground-based field robot	RGB	Semantic segmentation	Identification of downy mildew	OA	Downloadable zip file	–	https://ars.els-cdn.com/content/image/1-s2.0-S2352340921005345-mmc1.zip	JPEG	[137]
The MangoNet semantic dataset	India	*Mangifera indica*	Shoot	49[i]	Handheld visible light camera in the field	RGB	Semantic segmentation	Identification and counting of fruits	OA	Downloadable zip file	–	https://github.com/avadesh02/MangoNet-Semantic-Dataset	JPG	[138]

Leaves: India's most famous basil plant leaves quality dataset	India	Basil	Shoot	1131	Handheld visible light camera in the field	RGB	Image-level annotations	Identification and classification of biotic stress	OA	Download each image separately	https://dx.doi.org/10.21227/a4f6-4413	https://ieee-dataport.org/open-access/leaves-india%E2%80%99s-most-famous-basil-plant-leaves-quality-dataset	JPG	
A database of leaf images: practice towards plant conservation with plant pathology	Jammu, Kashmir	Multiple species[m]	A single detached leaf	4503	Handheld visible light camera in a controlled environment	RGB	Image-level annotations	Identification of biotic stress	OA	Downloadable zip file	http://doi.org/10.17632/hb74ynkjcn.4	https://data.mendeley.com/datasets/hb74ynkjcn/4	JPG	[139]
A citrus fruits and leaves dataset for detection and classification of citrus diseases through machine learning	Pakistan	Citrus	Shoot A single detached leaf	150 609	Handheld visible light camera in the field	RGB	Image-level annotations	Identification and classification of biotic stress	OA	Downloadable zip file	http://doi.org/10.17632/3f83gxmv57.2	https://data.mendeley.com/datasets/3f83gxmv57/2	JPG	[140, 141]
Supporting data for "morphometric analysis of Passiflora leaves: the relationship between landmarks of the vasculature and elliptical Fourier descriptors of the blade"	Brazil	40 *Passiflora* species[n]	A single detached leaf	5767	Multifunction printer	RGB	Point annotations of leaf edges	Classification of species	OA	Download each image separately	http://doi.org/10.5524/100251	http://gigadb.org/dataset/100251	TIF	[142, 143]
QuinceSet: dataset of annotated Japanese quince images for object detection	Latvia	*Chaenomeles japonica*	Shoot	1515	Handheld visible light camera in the field	RGB	Image-level annotations, bounding boxes	Identification and localization of fruits	OA	Downloadable zip file	https://doi.org/10.5281/zenodo.6402251	https://zenodo.org/record/6402251	JPG	[144]
Thermal images - diseased & healthy leaves	India	*Oryza sativa*	An intact leaf	636	Handheld TIR camera in the field	TIR	Image-level annotations	Classification of biotic stress	OA	Create Kaggle account to download zip file	–	https://www.kaggle.com/sujaradha/thermal-images-diseased-healthy-leaves-paddy	JPG	
Date fruit dataset	Saudi Arabia	*Phoenix dactylifera*	Shoot	8079	Handheld visible light camera in the	RGB	Image-level annotations	Detection and maturity classification of	OA	Create IEEE DataPort account to download zip	http://doi.org/10.21227/x46j-	https://ieee-dataport.org/open-access/date-	JPG	[145, 146]

			Fruit bunch	152	field			fruits		file	sk98	fruit-dataset-automated-harvesting-and-visual-yield-estimation		
			Date fruit	256										
Image set for deep learning: field images of maize annotated with disease symptoms	NY - USA	*Zea mays*	Shoot	7669	UAV	RGB	Line and spline annotation of lesions	Identification of northern leaf blight infected plants	OA	Downloadable zip file	–	https://osf.io/p67rz	JPG	[147]
				10533	Handheld visible light camera in the field									
Vegetable crops dataset for proximal sensing	France	*Z. mays*	Shoot	1065	Manned mobile platform	RGB	Image-level annotations, bounding boxes, point annotations	Identification and localization of crop stems	OA	Downloadable zip file	https://doi.org/10.17632/d7kbzjr83k.1	https://data.mendeley.com/datasets/d7kbzjr83k/1	JPG	[148]
		P. vulgaris		779										
		Allium ampeloprasum		601										
Pheno4D	Germany	Maize	Shoot	84	Controlled environment stationary platform	3D laser scanner	Semantic segmentation	Estimation of plant traits, growth analysis	OA	Downloadable zip file	–	https://www.ipb.uni-bonn.de/data/pheno4d	XYZ	[149]
		Tomato		140										
JMuBEN	Kenya	Arabica coffee	A single detached leaf	22591	Handheld visible light camera in the field	RGB	Image-level annotations	Identification and classification of biotic stress	OA	Downloadable zip file	https://doi.org/10.17632/t2r6rszp5c.1	https://data.mendeley.com/datasets/t2r6rszp5c/1	JPG	[150]
JMuBEN2				35964							https://doi.org/10.17632/tgv3zb82nd.1	https://data.mendeley.com/datasets/tgv3zb82nd/1		
BRACOL - A Brazilian Arabica coffee leaf images dataset	Brazil	Arabica coffee	A single detached leaf	1747	Handheld visible light camera in the field	RGB	Image-level annotations	Identification and classification of biotic stress	OA	Downloadable zip file	https://doi.org/10.17632/yy2k5y8mxg.1	https://data.mendeley.com/datasets/yy2k5y8mxg/1	JPG	[151]
RoCoLe: a robusta coffee leaf images dataset	Ecuador	*Robusta* coffee	An intact leaf	1560	Handheld visible light camera in the field	RGB	Image-level annotations	Identification and classification of biotic stress	OA	Downloadable zip file	https://doi.org/10.17632/c5yvn32dzg.2	https://data.mendeley.com/datasets/c5yvn32dzg/2	JPG	[152]
KOMATSUNA dataset for instance segmentation, tracking and reconstruction	Japan	*B. rapa*	Shoot	180	Controlled environment stationary platform	RGB	Semantic segmentation	Leaf segmentation and plant growth measurement	OA	Downloadable zip file	–	https://limu.ait.kyushu-u.ac.jp/~agri/komatsuna	PNG	[153]
				60		RGB-depth								
Single tree point clouds from terrestrial laser scanning	Germany, OR - USA	*Quercus petraea*	Shoot	22[k]	Terrestrial laser scanner	3D laser scanner	Image-level annotations	Classification of tree species	OA	Downloadable zip file	https://doi.org/10.25625/FOHUJM	https://data.goettingen-research-online.de/dataset.xhtml?persistentId=doi:10.25625/FOHUJM	XYZ[o]	[154]
		Fraxinus excelsior		39[k]										
		Picea abies		158										
		Pinus sylvestris		25[k]										
		Q. rubra		100										

		Fagus sylvatica		163										
		Pseudotzuga menziesii		183										
Early-crop-weed	Greece	*S. lycopersicum*	Shoot	202	Handheld visible light camera in the field	RGB	Image-level annotations	Identification of crops and weeds	OA	Downloadable zip file	–	https://github.com/AUAgroup/early-crop-weed	JPG	[155]
		Gossypium hirsutum		48[k]										
		S. nigrum		130										
		Abutilon theophrasti		124										
DeepWeeds	Australia	Multiple species[p]	Shoot	8403	Ground-based field robot	RGB	Image-level annotations	Classification of weed species	OA	Downloadable zip file	–	https://github.com/AlexOlsen/DeepWeeds	JPG	[156]

[a]Abbreviations: ACFR, Australian center for field robotics; CF, chlorophyll fluorescence; FGVC7/8, the seventh/eight workshop on fine-grained visual categorization; GARNICS, gardening with a cognitive system; HDF5, hierarchical data format version 5; HS, hyperspectral; JMuBEN, Jepkoech, Mugo and Benson; JPG, joint photographic experts group; LiDAR, light detection and ranging; MAT, Matlab; MS, multispectral; MSU-PID, Michigan State University-plant imagery database; NIR, near infrared; OA, open access; PID, persistent identifier; RGB, red–green–blue; TIF, tag image file format; TIR, thermal infrared; UAV, unmanned aerial vehicle.

[b]We identified 56 publicly available global datasets and their characteristics using Google's search engine and Google Dataset Search. The search combined terms describing various plant organ systems, sensors, artificial intelligence (AI) techniques, as well as dataset and database. All pages for each search were systematically collated and screened. Additional datasets are available in repositories containing large amounts of OA imaging data. Repositories such as the National Ecological Observatory Network (https://data.neonscience.org/data-products/explore), Leafsnap (http://leafsnap.com), the Institut National de la Recherche Agronomique (https://data.inrae.fr), the United States Geological Survey (https://www.usgs.gov/products/data-and-tools/science-datasets), the National Aeronautics and Space Administration earth science data (https://earthdata.nasa.gov), the plant genomics and phenomics research data repository (https://edal-pgp.ipk-gatersleben.de), the computer vision and biosystems signal processing group (https://vision.eng.au.dk/data-sets), the Transportation Energy Resources from Renewable Agriculture Phenotyping Reference Platform (https://terraref.ncsa.illinois.edu/clowder), figshare (https://figshare.com), Dryad (http://datadryad.org), the International Maize and Wheat Improvement Center (https://data.cimmyt.org) and the *Arabidopsis thaliana* phenotyping database (Phenopsis DB, http://bioweb.supagro.inra.fr/phenopsis) provide datasets in downloadable zip files. Similarly, the Oak Ridge National Laboratory Distributed Active Archive (https://daac.ornl.gov/get_data/#themes) provides datasets in downloadable zip files after registering for an account, as well as the University of Nebraska-Lincolin Plant Vision Initiative (https://plantvision.unl.edu/dataset) and the X-Plant (http://www.x-plant.org) after filling a form. Other sources such as the online database for plant image analysis software tools (https://www.quantitative-plant.org/dataset) and the registry of research data repositories (https://www.re3data.org) are designed specifically for the discovery of datasets in various repositories.

[c]Plant species were reported whenever they were available in the corresponding referenced paper(s); common names were reported otherwise.

[d]Images of datasets with semantic segmentation annotations are completely annotated images, where a class is assigned to each pixel.

[e]No datasets were excluded on the basis of access type (i.e., OA, data available on request, or OA with barriers – datasets fulfilling criteria for OA but being inaccessible because of unpredictable reasons such as broken hyperlinks).

[f]Data PID is a long-lasting digital reference to a dataset, such as a digital object identifier (DOI). A dash (–) indicates that no PIDs are available. DOIs for datasets can be issued automatically by the hosting repositories (e.g., Zenodo, GigaDB, Mendeley Data, and IEEE DataPorts). As datasets should be cited to ensure credit to those who produced and curated them, we recommend that they should include a PID and the minimum metadata suggested by DataCite (a non-profit membership organization that provides DOIs for research data) and FORCE11 (a community of scholars, librarians, archivists, publishers and research funders), i.e., author, year, title, and repository. Data producers can be inferred based on the author contributions of the corresponding referenced paper(s) while data curators can be inferred based on the author(s) that published the dataset to a repository.

[g]File format defines the structure and encoding of the data stored in it and thus guides researchers on how to programmatically input such data to their AI algorithms.

[h]The same number of images was taken with each sensor.

[i]Datasets containing more than one object per image (e.g., multiple hypocotyls, fruits, flowers). When segmented, each image could become hundreds of samples to train an AI algorithm.

[j]Two multispectral cameras were used: a five-band RedEdge-M camera in Germany and a four-band Sequoia camera in Switzerland.

[k]For datasets with small image number, transfer learning can be applied, giving an AI model a warm start by applying information learned from another previously trained model.

[l]Crops and their corresponding number of images: Apple, 3171; blueberry, 1502; cherry, 1906; corn, 3852; grapevine, 4062; orange, 5507; peach, 2657; pepper, 2475; potato, 2152; raspberry, 371; soybean, 6925; strawberry, 1565; tomato, 18160.

[m]Species and their corresponding number of images: *M. indica*, 435; *Terminalia Arjuna*, 452; *Alstonia Scholaris*, 433; *Psidium guajava*, 419; *Aegle marmelos*, 118; *Syzgium cumini*, 624; *Jatropha curcas*, 257; *Pongamia Pinnata*, 598; *Ocimum basilicum*, 149; *Punica granatum*, 559; *Platanus orientalis*, 223; *C. limon*, 236.

[n]*Passiflora* species and their corresponding number of images: *P. coriacea,* 208; *P. misera,* 215; *P. biflora,* 105; *P. capsularis,* 118; *P. micropetala,* 68; *P. organensis,* 84; *P. pohlii,* 16; *P. rubra,* 87; *P. tricuspis,* 257; *P. caerulea,* 99; *P. cincinnata,* 84; *P. edmundoi,* 111; *P. gibertii*, 192; *P. hatschbachii,* 132; *P. kermesina,* 113; *P. mollissima,* 69; *P. setacea,* 189; *P. suberosa,* 352; *P. tenuifila,* 113; *P. amethystina,* 119; *P. foetida,* 304; *P. gracilis,* 81; *P. morifolia,* 57; *P. actinia,* 95; *P. miersii,* 133; *P. sidifolia,* 145; *P. triloba,* 295; *P. alata,* 235; *P. edulis,* 119; *P. ligularis,* 139; *P. nitida,* 62; *P. racemosa,* 194; *P. villosa,* 58; *P. coccinea,* 169; *P. cristalina,* 220; *P. galbana,* 161; *P. malacophylla,* 168; *P. maliformis,* 156; *P. miniata,* 129; *P. mucronata,* 116.
[o]A point cloud data file in XYZ format contains rows of data, each consisting of x, y, and z coordinates of a point.
[p]Species and their corresponding number of images: *Ziziphus mauritiana*, 1125; *Lantana camara*, 1064; *Parkinsonia aculeata*, 1031; *Parthenium hysterophorus*, 1022; *Vachellia nilotica*, 1062; *Cryptostegia grandiflora*, 1009; *Chromolaena odorata*, 1074; *Stachytarpheta spp.*, 1016.

Glossary

Bias: systematic errors in the ability of AI models to make correct predictions.

Compute node: a backend node used for computing in a cluster and reached via a frontend node.

Computer cluster: a group of interconnected computers working together as a single, integrated computing resource.

Confusion matrix: a visual representation that describes the complete performance of an AI model, summarizing its predictions in four categories: true-positives, true-negatives, false-positives, and false-negatives.

Crowdsourced: the act of collecting data by soliciting contributions from a large group of people rather than from traditional experiments.

Explicit knowledge: the human knowledge that can be readily assembled and passed on by written or verbal instruction. Metadata is explicit knowledge about data.

Federated learning (FL): a collaborative AI training paradigm in which copies of a model are distributed to devices, where data is stored, for local training, and the resulting model weights, rather than the data, are sent back to a central server to update the main model.

GPU-accelerated: a backend node used mainly for accelerating computing, connected in a heterogeneous matter in a computer cluster.

Human-in-the-loop (HITL): an approach that aims to achieve what neither a human nor a machine can do on their own; it leverages a continuous feedback loop between them to train, evaluate, and deploy AI models that continuously learn and improve their prediction accuracy.

Hyperparameters: a group of variables whose values cannot be estimated from data and are manually tweaked to determine the optimal configuration to train a specific model (e.g., learning rate, batch size, number of training epochs).

Notebook: a web browser-based interactive computing environment that can be used to combine software code, computational output, explanatory text, and multimedia resources in a single document.

Parallel computing: a form of computation in which multiple compute nodes operating simultaneously are used to solve a large problem broken into independent smaller parts that can be processed concurrently.

Quantum bit (qubit): the quantum analogue of a classical bit; it may adopt the states 0, 1, or any possible combination of both states.

Quantum processing unit (QPU): a computational unit that leverages quantum mechanical phenomena to manipulate information, relying on qubits.

Quantum volume (QV): a metric that measures the performance of a quantum computer taking into account its number of qubits and error rates.

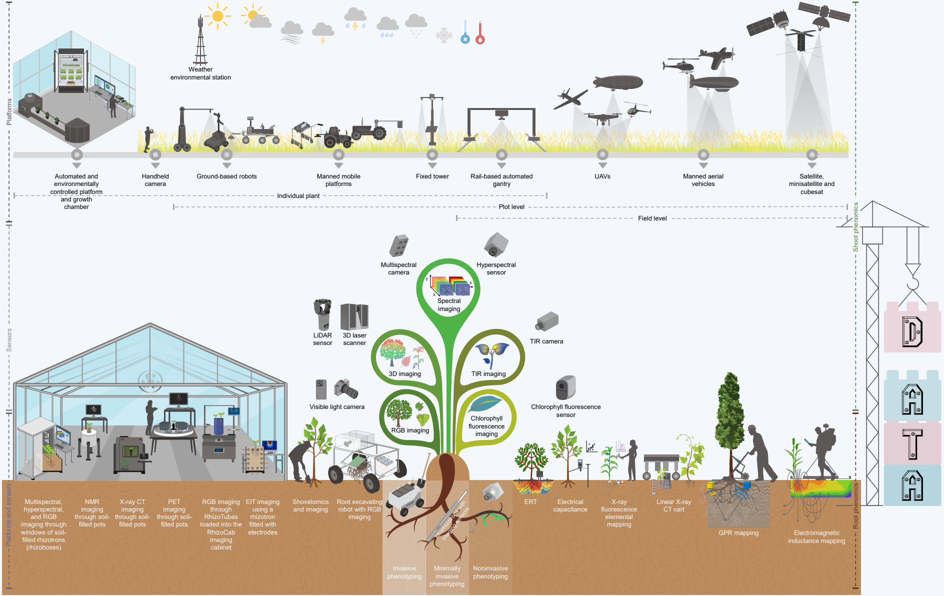
Synthetic data: data generated artificially using AI algorithms when real data cannot be collected in sufficient amounts.

Tacit knowledge: the know-how, skills, and intuition that live in the individual's experiences and are hard to impart or transfer to others. It can be shared through advances in information and communications technology, and thus becomes explicit.

Transfer learning (TL): a technique in which an AI algorithm reuses parts of a previously trained model on a new model to perform a different but similar task.

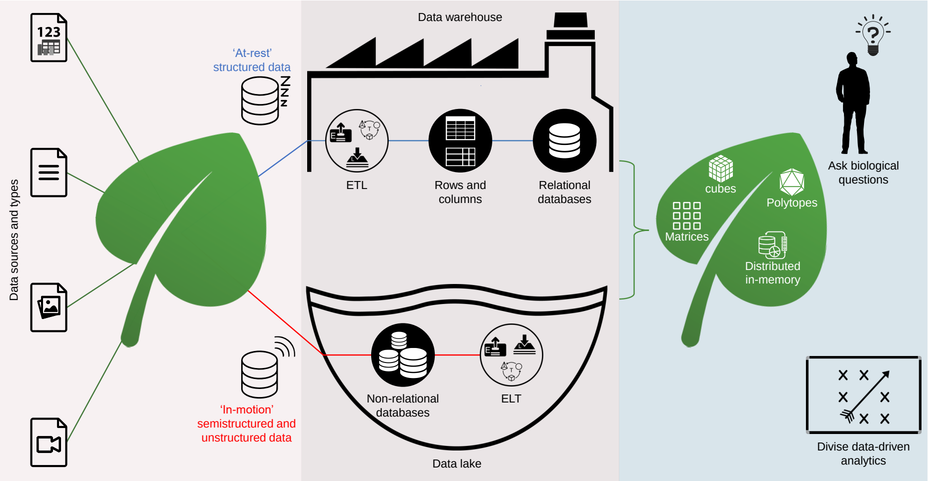
Trustworthiness: a quality of an AI model working reliably in ways that anyone can trust; it should be (i) lawful, ensuring compliance with all applicable laws and regulations; (ii) ethical, demonstrating adherence to ethical principles and values, (iii) robust, able to deal with bias during all of its lifecycle; and (iv) explainable.

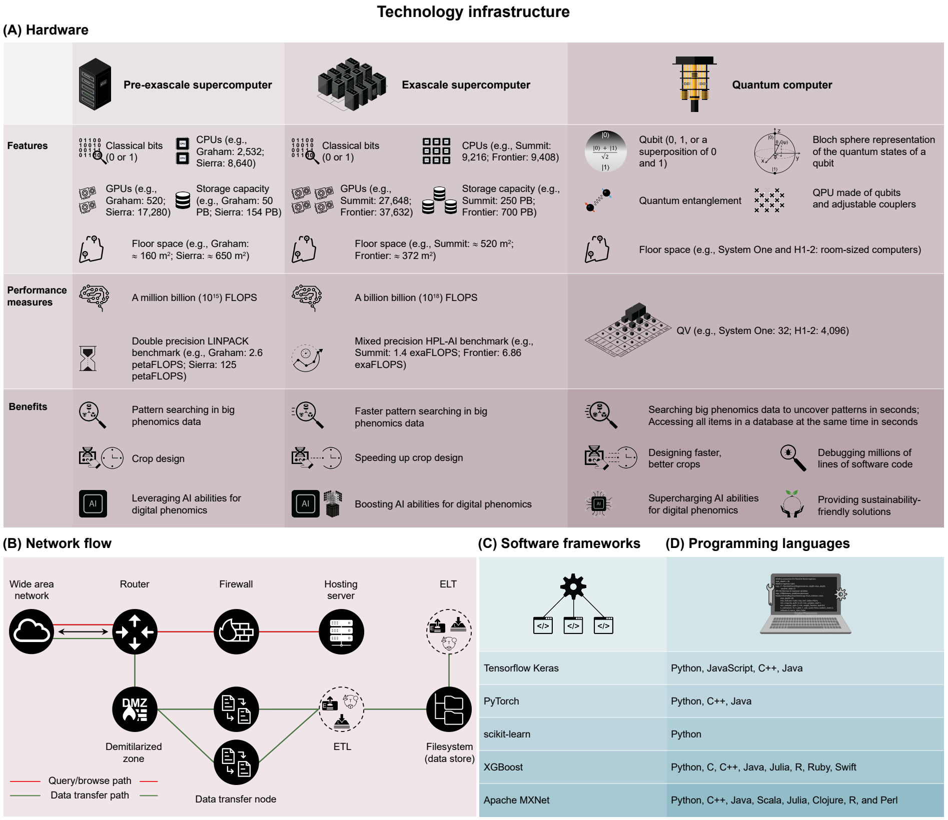
Unbalanced dataset: a dataset having certain classes contain substantially more training examples than other classes, misleading the classifier algorithm to overlearn the majority classes and to perform poorly in the prediction of the minority classes.

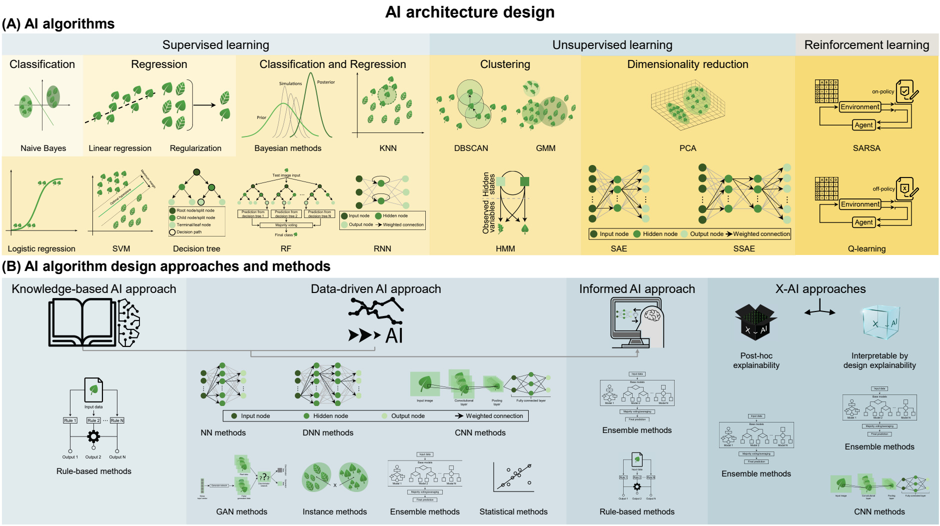


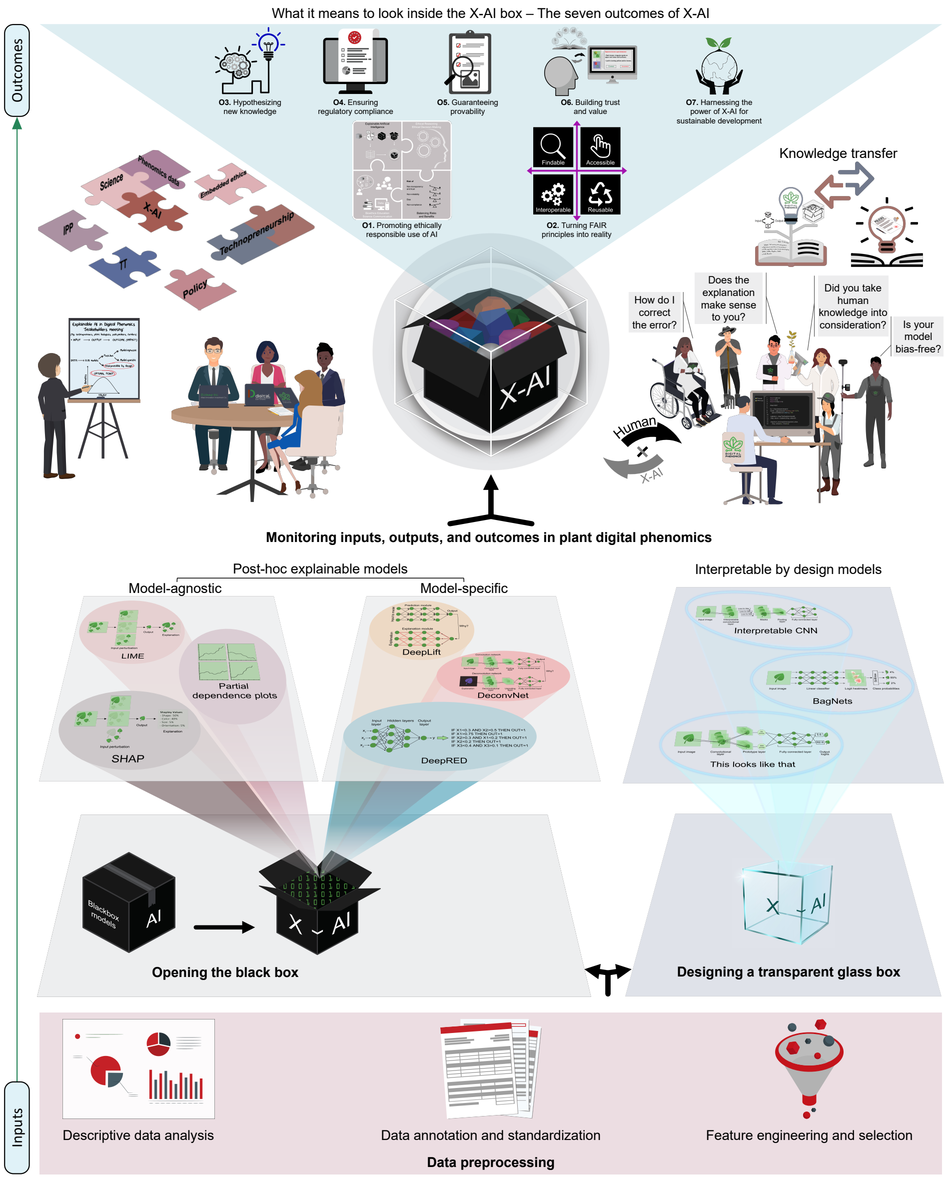
Platforms
Weather environmental station
Automated and environmentally controlled platform and growth chamber
Handheld camera
Ground-based robots
Manned mobile platforms
Fixed tower
Rail-based automated gantry
UAVs
Manned aerial vehicles
Satellite, minisatellite and cubesat
Individual plant
Plot level
Field level
Shoot phenomics
Sensors
Multispectral camera
Hyperspectral sensor
Spectral imaging
LiDAR sensor
3D laser scanner
TIR camera
3D imaging
TIR imaging
Visible light camera
Chlorophyll fluorescence sensor
RGB imaging
Chlorophyll fluorescence imaging
Platforms and sensors
Multispectral, hyperspectral, and RGB imaging through windows of soil-filled rhizotrons (rhizoboxes)
NMR imaging through soil-filled pots
X-ray CT imaging through soil-filled pots
PET imaging through soil-filled pots
RGB imaging through RhizoTubes loaded into the RhizoCab imaging cabinet
EIT imaging using a rhizotron fitted with electrodes
Shovelomics and imaging
Root excavating robot with RGB imaging
Minirhizotrons with imaging sensors
ERT
Electrical capacitance
X-ray fluorescence elemental mapping
Linear X-ray CT cart
GPR mapping
Electromagnetic inductance mapping
Invasive phenotyping
Minimally invasive phenotyping
Noninvasive phenotyping
Root phenomics
D
A
T
A

Data architecture

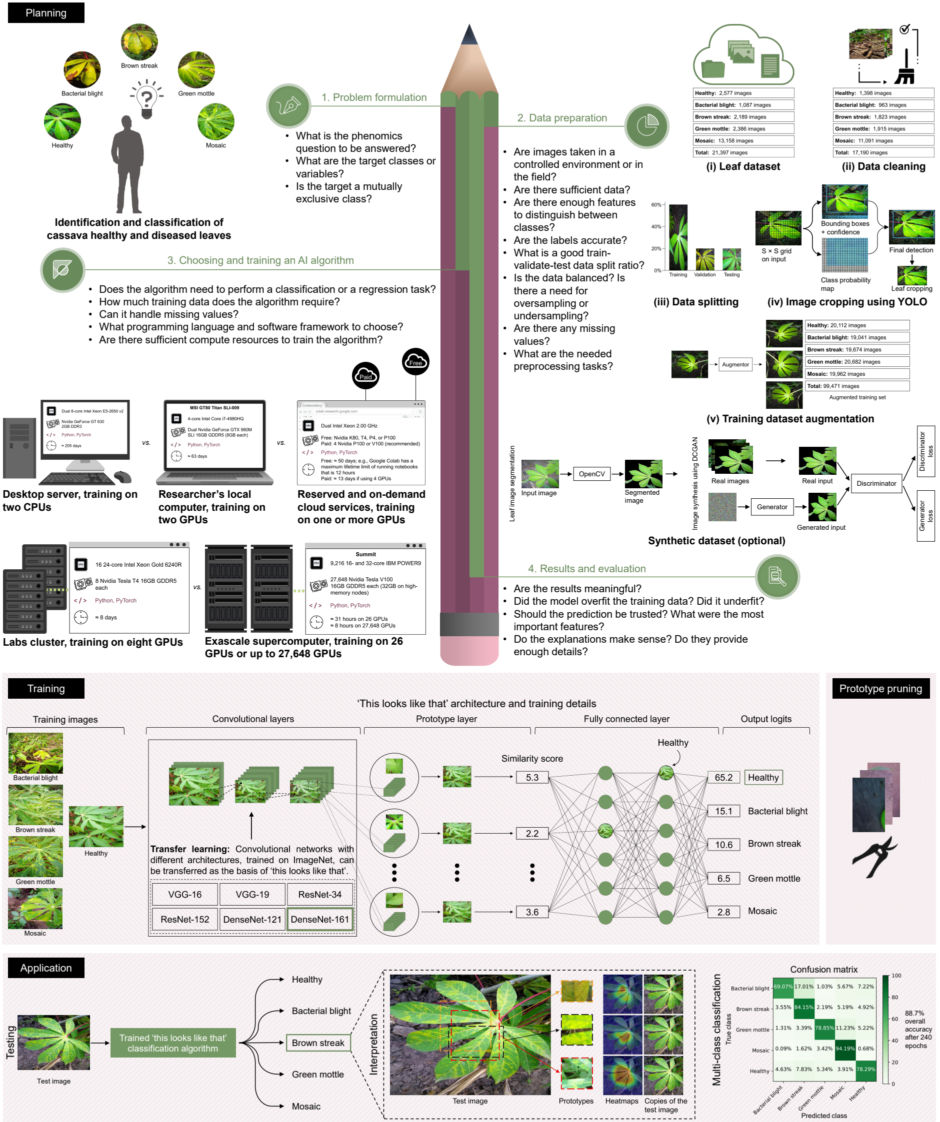



Technology infrastructure
(A) Hardware
Pre-exascale supercomputer
Exascale supercomputer
Quantum computer
Features
Classical bits (0 or 1)
CPUs (e.g., Graham: 2,532; Sierra: 8,640)
GPUs (e.g., Graham: 520; Sierra: 17,280)
Storage capacity (e.g., Graham: 50 PB; Sierra: 154 PB)
Floor space (e.g., Graham: ≈ 160 m²; Sierra: ≈ 650 m²)
Classical bits (0 or 1)
CPUs (e.g., Summit: 9,216; Frontier: 9,408)
GPUs (e.g., Summit: 27,648; Frontier: 37,632)
Storage capacity (e.g., Summit: 250 PB; Frontier: 700 PB)
Floor space (e.g., Summit: ≈ 520 m²; Frontier: ≈ 372 m²)
Qubit (0, 1, or a superposition of 0 and 1)
Bloch sphere representation of the quantum states of a qubit
Quantum entanglement
QPU made of qubits and adjustable couplers
Floor space (e.g., System One and H1-2: room-sized computers)
Performance measures
A million billion (10¹⁵) FLOPS
Double precision LINPACK benchmark (e.g., Graham: 2.6 petaFLOPS; Sierra: 125 petaFLOPS)
A billion billion (10¹⁸) FLOPS
Mixed precision HPL-AI benchmark (e.g., Summit: 1.4 exaFLOPS; Frontier: 6.86 exaFLOPS)
QV (e.g., System One: 32; H1-2: 4,096)
Benefits
Pattern searching in big phenomics data
Crop design
Leveraging AI abilities for digital phenomics
Faster pattern searching in big phenomics data
Speeding up crop design
Boosting AI abilities for digital phenomics
Searching big phenomics data to uncover patterns in seconds; Accessing all items in a database at the same time in seconds
Designing faster, better crops
Debugging millions of lines of software code
Supercharging AI abilities for digital phenomics
Providing sustainability-friendly solutions
(B) Network flow
Wide area network
Router
Firewall
Hosting server
ELT
DMZ
Demitilarized zone
ETL
Filesystem (data store)
Data transfer node
Query/browse path
Data transfer path
(C) Software frameworks
(D) Programming languages
Tensorflow Keras
Python, JavaScript, C++, Java
PyTorch
Python, C++, Java
scikit-learn
Python
XGBoost
Python, C, C++, Java, Julia, R, Ruby, Swift
Apache MXNet
Python, C++, Java, Scala, Julia, Clojure, R, and Perl


AI architecture design
(A) AI algorithms
Supervised learning
Unsupervised learning
Reinforcement learning
Classification
Regression
Classification and Regression
Clustering
Dimensionality reduction
Naive Bayes
Linear regression
Regularization
Bayesian methods
KNN
DBSCAN
GMM
PCA
SARSA
Logistic regression
SVM
Decision tree
RF
RNN
HMM
SAE
SSAE
Q-learning
Input node
Hidden node
Output node
Weighted connection
on-policy
off-policy
Environment
Agent
(B) AI algorithm design approaches and methods
Knowledge-based AI approach
Data-driven AI approach
Informed AI approach
X-AI approaches
Post-hoc explainability
Interpretable by design explainability
Rule-based methods
NN methods
DNN methods
CNN methods
Ensemble methods
GAN methods
Instance methods
Statistical methods



What it means to look inside the X-AI box – The seven outcomes of X-AI
Outcomes
O3. Hypothesizing new knowledge
O4. Ensuring regulatory compliance
O5. Guaranteeing provability
O6. Building trust and value
O7. Harnessing the power of X-AI for sustainable development
O1. Promoting ethically responsible use of AI
Findable
Accessible
Interoperable
Reusable
O2. Turning FAIR principles into reality
Knowledge transfer
Science
Phenomics data
Embedded ethics
X-AI
IPP
Technopreneurship
TT
Policy
How do I correct the error?
Does the explanation make sense to you?
Did you take human knowledge into consideration?
Is your model bias-free?
Human
X-AI
Monitoring inputs, outputs, and outcomes in plant digital phenomics
Post-hoc explainable models
Model-agnostic
Model-specific
Interpretable by design models
LIME
Partial dependence plots
SHAP
DeepLift
DeconvNet
DeepRED
Interpretable CNN
BagNets
This looks like that
Blackbox models
AI
X AI
Opening the black box
Designing a transparent glass box
Descriptive data analysis
Data annotation and standardization
Feature engineering and selection
Data preprocessing
Inputs



Planning
Brown streak
Bacterial blight
Green mottle
Healthy
Mosaic
Identification and classification of cassava healthy and diseased leaves
1. Problem formulation
• What is the phenomics question to be answered?
• What are the target classes or variables?
• Is the target a mutually exclusive class?
2. Data preparation
• Are images taken in a controlled environment or in the field?
• Are there sufficient data?
• Are there enough features to distinguish between classes?
• Are the labels accurate?
• What is a good train-validate-test data split ratio?
• Is the data balanced? Is there a need for oversampling or undersampling?
• Are there any missing values?
• What are the needed preprocessing tasks?
Healthy: 2,577 images
Bacterial blight: 1,087 images
Brown streak: 2,189 images
Green mottle: 2,386 images
Mosaic: 13,158 images
Total: 21,397 images
(i) Leaf dataset
Healthy: 1,398 images
Bacterial blight: 963 images
Brown streak: 1,823 images
Green mottle: 1,915 images
Mosaic: 11,091 images
Total: 17,190 images
(ii) Data cleaning
60%
40%
20%
0%
Training
Validation
Testing
(iii) Data splitting
S × S grid on input
Bounding boxes + confidence
Class probability map
Final detection
Leaf cropping
(iv) Image cropping using YOLO
Augmentor
Healthy: 20,112 images
Bacterial blight: 19,041 images
Brown streak: 19,674 images
Green mottle: 20,682 images
Mosaic: 19,962 images
Total: 99,471 images
Augmented training set
(v) Training dataset augmentation
Leaf image segmentation
Input image
OpenCV
Segmented image
Image synthesis using DCGAN
Real images
Real input
Discriminator
Discriminator loss
Generator
Generated input
Generator loss
Synthetic dataset (optional)
3. Choosing and training an AI algorithm
• Does the algorithm need to perform a classification or a regression task?
• How much training data does the algorithm require?
• Can it handle missing values?
• What programming language and software framework to choose?
• Are there sufficient compute resources to train the algorithm?
Paid
Free
Dual 8-core Intel Xeon E5-2650 v2
Nvidia GeForce GT 630 2GB DDR3
Python, PyTorch
≈ 205 days
vs.
MSI GT80 Titan SLI-009
4-core Intel Core i7-4980HQ
Dual Nvidia GeForce GTX 980M SLI 16GB GDDR5 (8GB each)
Python, PyTorch
≈ 63 days
vs.
Colaboratory
colab.research.google.com
Dual Intel Xeon 2.00 GHz
Free: Nvidia K80, T4, P4, or P100
Paid: 4 Nvidia P100 or V100 (recommended)
Python, PyTorch
Free: ≈ 50 days; e.g., Google Colab has a maximum lifetime limit of running notebooks that is 12 hours
Paid: ≈ 13 days if using 4 GPUs
Desktop server, training on two CPUs
Researcher's local computer, training on two GPUs
Reserved and on-demand cloud services, training on one or more GPUs
16 24-core Intel Xeon Gold 6240R
8 Nvidia Tesla T4 16GB GDDR5 each
Python, PyTorch
≈ 8 days
vs.
Summit
9,216 16- and 32-core IBM POWER9
27,648 Nvidia Tesla V100 16GB GDDR5 each (32GB on high-memory nodes)
Python, PyTorch
≈ 31 hours on 26 GPUs
≈ 8 hours on 27,648 GPUs
Labs cluster, training on eight GPUs
Exascale supercomputer, training on 26 GPUs or up to 27,648 GPUs
4. Results and evaluation
• Are the results meaningful?
• Did the model overfit the training data? Did it underfit?
• Should the prediction be trusted? What were the most important features?
• Do the explanations make sense? Do they provide enough details?
Training
'This looks like that' architecture and training details
Training images
Convolutional layers
Prototype layer
Fully connected layer
Output logits
Bacterial blight
Brown streak
Healthy
Green mottle
Mosaic
Transfer learning: Convolutional networks with different architectures, trained on ImageNet, can be transferred as the basis of 'this looks like that'.
VGG-16
VGG-19
ResNet-34
ResNet-152
DenseNet-121
DenseNet-161
Similarity score
5.3
2.2
3.6
Healthy
65.2
Healthy
15.1
Bacterial blight
10.6
Brown streak
6.5
Green mottle
2.8
Mosaic
Prototype pruning
Application
Testing
Test image
Trained 'this looks like that' classification algorithm
Healthy
Bacterial blight
Brown streak
Green mottle
Mosaic
Interpretation
Test image
Prototypes
Heatmaps
Copies of the test image
Multi-class classification
Confusion matrix
True class
Bacterial blight 69.07% 17.01% 1.03% 5.67% 7.22%
Brown streak 3.55% 84.15% 2.19% 5.19% 4.92%
Green mottle 1.31% 3.39% 78.85% 11.23% 5.22%
Mosaic 0.09% 1.62% 3.42% 94.19% 0.68%
Healthy 4.63% 7.83% 5.34% 3.91% 78.29%
Bacterial blight
Brown streak
Green mottle
Mosaic
Healthy
Predicted class
100
80
60
40
20
0
88.7% overall accuracy after 240 epochs