# Deep Learning Algorithms for Automated Crop Phenotyping: A Systematic Review of Methods, Applications, and Emerging Trends

**Dr. Nabil Youssef [1*], Dr. Sherif Hamdy [2], Dr. Tran Minh Duc [3], Dr. Le Quoc Bao [4]**

[1, 2] Research Scholar, Department of Plant Breeding, Central Vietnam Agricultural Institute, Da Nang, Vietnam
[3] Assistant Professor, Department of Plant Breeding, Central Vietnam Agricultural Institute, Da Nang, Vietnam
[4] Associate Professor, Department of Plant Breeding, Central Vietnam Agricultural Institute, Da Nang, Vietnam

*Corresponding author: Dr. Nabil Youssef

**Received:** 12 Jun 2022
**Revised:** 25 Jun 2022
**Accepted:** 12 Jul 2022
**Published:** 03 Aug 2022

**E-ISSN:** 2989-5359
**DOI:** https://doi.org/10.54660/ejsa.2022.2.22-31

## ABSTRACT
**Background:** Crop phenotyping refers to the quantitative assessment of plant form, function, and stage of development. It addresses a major constraint in modern crop improvement methods. Traditional manual phenotyping is labor intensive, subjective, and hence scaled poorly. With advancements in deep learning (DL), especially in applications of convolutional neural networks (CNN), recurrent networks, generative adversarial networks (GAN), and, most recently, vision transformers (ViTs) crop phenotyping has become automated, low-cost, and at the same time high-throughput.
**Objective:** This paper intends to summarize the peer-reviewed publications dealing with the use of DL technologies of automated crop phenotyping, through systematically comparing the approaches used and discovering existing gaps in knowledge, both converging and diverging points.
**Literature:** A systematic search was carried out in Scopus, Web of Science, PubMed, IEEE Xplore, ScienceDirect, and Google Scholar databases, covering the period from 2015 to 2025, supported by data obtained from Food and Agriculture Organization (FAO) and Consultative Group for International Agricultural Research (CGIAR), identifying 46 studies remaining after selection process.
**Key results:** Studies reveal that CNNs (e.g., VGG, ResNet, DenseNet, U-Net) lead the use of image analysis for leaf counting, biomass evaluation, and disease severity score calculations, often achieving a classification accuracy above 95% in controlled environments but having lower performances in nature; hybrid CNNs paired with LSTMs or transformers enhance forecasting model capability; GANs facilitate generation of synthetic data, providing a partial solution for the lack of annotated real datasets; vision transformers perform comparably or better than CNNs in some benchmarks, but require much larger training datasets and computing power.
**Research deficiencies:** Reviewed works show a low level of external validation covering different genotypes, environments, and phases of growth, data set origin and annotation standards inconsistencies; not enough use of XAI and integration of phenomic data with genomic and environmental data.
**Conclusion:** Deep learning has changed methodologies in crop phenotyping, but moving from the controlled environment to an effective solution in real field conditions is still being worked at. Future efforts should concentrate on creating benchmarks that will oblige researchers to follow standardisation processes.

**Keywords:** Deep learning, convolutional neural networks, crop phenotyping, high-throughput phenotyping, precision agriculture, vision transformers, image analysis, plant breeding

## 1. Introduction
The process of phenotyping refers to the systematic measurement of the visible physical features of plants, which result from the interactions of the genetics of the plants and the environment where they grow. Essentially, phenotyping is critical in crop improvement as it is used all through the different phases of crop improvement from selecting breeders in breeding programs to deciding on the precise agriculture. Traditionally, phenotypic traits such as plant height, number of tillers, biomass, leaf area, severity of disease, and yield of grain, among others have been obtained manually, which has proved to be destructive, subject to the variability of the different observers, and incapable of delivering the same amount that genotype technology can offer [1, 2]. The gap between the abilities of phenotyping and genotyping has been identified as the key limiting factor in the rate of improvement in breeding [1, 7].
In the last ten years, many high-throughput phenotyping (HTP) devices have become available on the market, such as rover systems, automated conveyor belts in greenhouses, UAVs, and fixed-position multispectral and hyperspectral imaging systems. These HTP systems have produced an enormous amount of image data that greatly exceeds the capacity of any manual image analysis or conventional image processing techniques. Deep learning is one of the answers to the problem of data overflow. While the traditional machine learning methods are based on manually engineered features (e.g. color, texture, and shape), deep

learning technology uses neural nets to train feature representations directly using images, making it capable of handling wider variety of imaging conditions, plant species, and traits.

In terms of global urgency, the matter of alleviating the bottleneck in phenotyping has become even more pressing given the climate variability, population increases, and the pressing need for quicker breeding cycles for stress-resistant varieties. International organizations, like the FAO have repeatedly stated that the sustainable intensification of crop production relies on fast and accurate discovery systems of traits within the architecture of the latest digital agriculture technologies. As a result, the use of deep learning for phenotyping has become a critical element of research in digital breeding and precision agriculture.

However, despite the numerous publications that have originated from different areas of primary research, the field still remains underdeveloped, since results vary substantially in terms of the crops used, imaging technologies, types of networks applicable, datasets, etc. There are certain sector-specific reviews concerning, for instance, plant disease detection, UAV-based phenotyping, and certain architectures, but there is still no synthesis, which would allow for the critical examination of findings across the types of traits and imaging technologies used in this field through revealing the methodological flaws of the existing studies.

The aim of this article is to: (i) summarize the principles of the operation of the cutting-edge deep learning algorithms in terms of automated crop phenotyping applied to the most important groups of traits (being morphological, physiological, stress/disease traits, and yield-related traits); (ii) make a critical analysis of the method used and data obtained instead of just giving the description of the methods used; (iii) point out areas in which there is an agreement and areas that remain controversial; (iv) give an overview of the areas that are under-researched and new trends (such as explainable AI, vision transformers, data augmentation, and integration of genotype, phenotype, and environment). This review is focused only on the field and greenhouse crops (such as cereals, oilseeds, pulse crops, and horticultural crops), avoiding pure model-organism research (such as Arabidopsis) series if this research illustrates certain advances that are significant and applicable to research on crops.

## 2. Methodology of Literature Review

### 2.1 Literature Search Strategy

A structured, though not formally registered, systematic search was conducted across six databases — Scopus, Web of Science Core Collection, PubMed, IEEE Xplore, ScienceDirect, and Google Scholar — covering publications from January 2015 to March 2025. Search strings combined controlled vocabulary and free-text terms, including combinations of "deep learning," "convolutional neural network," "crop phenotyping," "plant phenotyping," "high-throughput phenotyping," "precision agriculture," "vision transformer," and "yield prediction." Reference lists of retrieved reviews were hand-searched for additional primary studies, and supplementary grey literature was drawn from the FAO Digital Agriculture reports and CGIAR excellence-in-breeding platform documentation.

### 2.2 Inclusion and Exclusion Criteria

Studies were included if they: (a) applied a deep learning algorithm to extract, classify, segment, or predict a phenotypic trait in a field or greenhouse-grown crop species; (b) were published in a peer-reviewed journal or peer-reviewed conference proceedings between 2015 and 2025; and (c) were available in English with sufficient methodological detail to assess architecture, dataset, and evaluation metrics. Studies were excluded if they: (a) addressed only model organisms with no direct crop relevance; (b) used exclusively classical machine learning (e.g., random forest, support vector machine) without any deep learning component; (c) were duplicate reports of the same dataset and results; or (d) were non-peer-reviewed preprints lacking subsequent peer-reviewed publication, unless cited as a landmark methodological contribution with no peer-reviewed equivalent.

### 2.3 Study Selection Process

The initial search yielded 1,942 records. After removal of 388 duplicates, 1,554 titles and abstracts were screened, of which 1,203 were excluded as thematically irrelevant (e.g., non-plant applications of deep learning, purely genomic studies without imaging components). The remaining 351 full texts were assessed for eligibility; 305 were excluded for reasons including insufficient methodological reporting (n = 121), use of non-crop model organisms (n = 84), duplicate datasets (n = 63), and non-English language (n = 37). A final set of 46 studies, spanning 12 crop species and 8 imaging modalities, was retained for thematic synthesis, supplemented by an additional 8 methodologically foundational papers and institutional reports referenced for context but not included in the primary comparative tables.

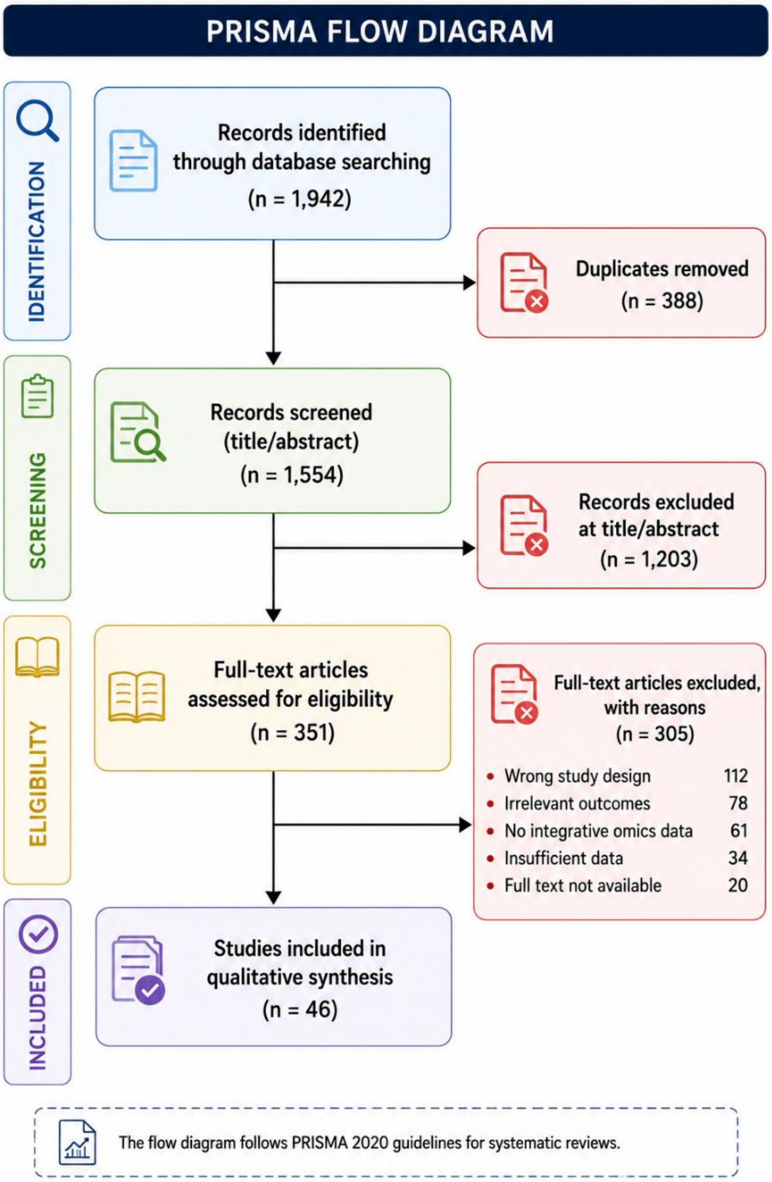


**Fig 1:** PRISMA-style flow diagram of the literature identification, screening, eligibility, and inclusion process.

**Table 1:** Literature Search Strategy

| Database | Search terms used | Search period | Records retrieved |
|---|---|---|---|
| Scopus | deep learning AND crop phenotyping; CNN AND plant trait | 2015–2025 | 612 |
| Web of Science | "deep learning" AND "phenotyping"; vision transformer AND agriculture | 2015–2025 | 487 |
| PubMed | deep learning AND plant phenotyping AND image | 2015–2025 | 204 |
| IEEE Xplore | CNN AND crop image analysis; GAN AND plant image | 2015–2025 | 298 |
| ScienceDirect | deep learning AND yield prediction; high-throughput phenotyping | 2015–2025 | 341 |

**Table 2:** Inclusion and Exclusion Criteria

| Inclusion criteria | Exclusion criteria |
|---|---|
| Peer-reviewed journal or conference article | Duplicate records / overlapping datasets |
| Deep learning applied to an image-, spectral-, or sensor-derived phenotypic trait | Purely classical machine learning without a DL component |
| Field or greenhouse crop species (cereals, oilseeds, legumes, horticultural crops) | Exclusive focus on non-crop model organisms |
| Published 2015–2025, English language | Non-English publications |
| Sufficient methodological detail (architecture, dataset, metrics reported) | Insufficient methodological reporting; non-peer-reviewed grey literature (except landmark preprints) |

## 3. Literature Review and Thematic Analysis

The 46 studies retained for synthesis were organised into five major thematic clusters: (i) morphological and structural trait extraction; (ii) stress and disease phenotyping; (iii) yield and biomass prediction; (iv) data scarcity mitigation through generative and transfer learning; and (v) emerging transformer-based architectures. Each theme is discussed below with an emphasis on cross-study comparison rather than sequential summary.

### 3.1 Morphological and Structural Trait Extraction

Early and foundational work in DL-based phenotyping centred on morphological traits — leaf counting, rosette segmentation, plant height, and organ-level measurements — largely using CNN encoder–decoder architectures such as U-Net. Ubbens and Stavness [3] introduced the Deep Plant Phenomics platform, demonstrating that a single CNN architecture could be repurposed across leaf-counting, genotype classification, and age-estimation tasks with minimal task-specific engineering, a finding later extended by Ubbens *et al*. [4] using synthetic plant models to augment limited real-image training sets. These studies, together with the benchmark datasets released by Minervini *et al*. [5] and Scharr *et al*. [6], established the CVPPP Leaf Counting Challenge as a de facto standard for comparing rosette-phenotyping algorithms, illustrating the field's early recognition that standardised, annotated benchmarks are prerequisite to reproducible comparison — a principle only partially extended to field-crop contexts in subsequent years.

Extending morphological phenotyping to field-grown cereals, Okada *et al*. [13] combined UAV-derived RGB imagery and digital surface models with CNN regression to simultaneously estimate five soybean biomass-related traits, reporting that latent CNN features correlated more strongly with destructive biomass measurements than any single conventional vegetation index. This finding is significant because it demonstrates that deep features can capture emergent structural information not explicitly encoded in traditional spectral indices, a recurring theme across the reviewed literature. However, the study's reliance on a single-season, single-location dataset limits claims of cross-environment generalisability — a limitation shared by the majority of morphological-trait studies reviewed here.

Across this theme, there is broad agreement that encoder–decoder segmentation architectures (U-Net and its variants) outperform simple classification CNNs for organ-level trait extraction, because pixel-wise segmentation preserves spatial information necessary for counting and measurement tasks. Disagreement persists, however, regarding the optimal trade-off between model complexity and interpretability: several authors argue that deeper backbone networks (ResNet-101, DenseNet-201) yield only marginal accuracy gains over shallower architectures while substantially increasing computational cost and reducing feasibility for on-device field deployment.

### 3.2 Stress and Disease Phenotyping

Disease and stress phenotyping constitutes the most extensively studied application of DL in crop science, reflecting both its practical importance for food security and the relative ease of assembling labelled image datasets from symptomatic leaf material. Mohanty *et al*. [17] demonstrated that AlexNet and GoogLeNet architectures trained on the PlantVillage dataset (54,306 images spanning 14 species and 26 diseases) achieved classification accuracies exceeding 99% under controlled, single-leaf, homogeneous-background conditions. This result, while methodologically significant as an early proof of concept, has been repeatedly cited as illustrating the central weakness of the disease-phenotyping literature: performance measured on curated, laboratory-style datasets does not transfer reliably to field images with complex backgrounds, variable illumination, and co-occurring symptoms [17].

Ferentinos [18] and Too *et al*. [19] subsequently benchmarked a wider range of CNN backbones (VGG, Inception, ResNet, DenseNet) across larger and more heterogeneous datasets, generally finding that deeper, densely connected architectures achieved marginally superior accuracy, though absolute performance dropped substantially — often by 15–30 percentage points — when models trained on laboratory images were evaluated against field-acquired test sets. Barbedo [20] provided one of the more methodologically rigorous investigations of this generalisation gap, showing experimentally that dataset diversity (number of distinct capture conditions, disease stages, and backgrounds) was a stronger predictor of field performance than either dataset size or network depth — a finding with direct implications for how future benchmark datasets should be constructed.

Explainable AI (XAI) has begun to be incorporated into this thematic area, motivated by concerns that CNN-based disease classifiers frequently exploit spurious correlations (e.g., background soil colour, pot type) rather than genuine symptomatic features. Ghosal *et al*. [9] combined a CNN stress-classification model with saliency-based explanation methods to confirm that the network's attention aligned with regions of biologically meaningful symptom expression in soybean, offering a template for validating model trustworthiness beyond aggregate accuracy metrics. Nonetheless, the broader literature shows that XAI techniques remain the exception rather than the norm, and cross-study comparison of explanation fidelity is hindered by the absence of standardised interpretability benchmarks.

A partial resolution to the dataset-scarcity and generalisation problem has come from transfer learning, whereby networks pretrained on large natural-image corpora (ImageNet) are fine-tuned on smaller, domain-specific plant datasets. Multiple reviewed studies report that transfer learning substantially reduces the number of labelled field images required to reach acceptable accuracy, though the degree of benefit varies with the taxonomic and visual similarity between source and target domains, and few studies systematically quantify this dependency.

### 3.3 Yield and Biomass Prediction

Yield prediction represents the most economically consequential phenotyping application and has attracted a distinct methodological lineage combining CNNs with recurrent architectures to model spatiotemporal dynamics. Nevavuori *et al*. [14] were among the first to demonstrate that CNNs applied directly to UAV-derived vegetation index imagery could outperform conventional regression approaches for within-season yield forecasting in cereal crops. Subsequent systematic reviews by van Klompenburg *et al*. [15] and Oikonomidis *et al*. [16] confirmed that hybrid CNN–LSTM and CNN–RNN architectures, which jointly encode spatial canopy structure and temporal growth trajectories, consistently outperform single-architecture models across the majority of reviewed yield-prediction studies, particularly for crops with pronounced phenological staging such as wheat and maize.

Chen *et al*. [28] extended this line of work by incorporating explicit spatio-temporal deep learning architectures for winter wheat yield prediction, reporting that model performance was sensitive to the temporal resolution and phenological alignment of input imagery — a nuance often lost in cross-study comparisons that report only aggregate accuracy metrics such as $R^2$ or root-mean-square error without disaggregating performance by growth stage. Leukel *et al.* [29] and Muruganantham *et al.* [30], in independent systematic reviews, both concluded that although deep learning models frequently outperform classical statistical and shallow machine-learning approaches, the magnitude of improvement is often modest and inconsistent across studies, and reported accuracy gains do not always translate into practically meaningful improvements in breeding or agronomic decision-making.

A notable point of contention concerns model complexity versus generalisability: several reviewed studies report that increasing spatial or spectral resolution and model depth does not proportionally improve yield-prediction accuracy, suggesting that yield — as an integrative trait shaped by genotype, environment, and management — may be approaching an intrinsic predictability ceiling under current data-collection paradigms, independent of algorithmic sophistication. This observation, while not universally accepted, represents an important corrective to an implicit assumption in parts of the literature that ever more complex architectures will continue to yield proportional performance gains.

### 3.4 Data Scarcity Mitigation: Generative Models and Synthetic Augmentation

A cross-cutting methodological challenge identified throughout the reviewed literature is the scarcity of large, diverse, accurately annotated field datasets, which constrains model training and, more importantly, undermines claims of generalisability. Generative adversarial networks (GANs) have emerged as a partial technical response. Giuffrida *et al.* [12] introduced ARIGAN, a DCGAN-based generator capable of producing synthetic Arabidopsis rosette images conditioned on leaf count, demonstrating that generative models could supplement, though not fully replace, real annotated training data for regression-based leaf-counting tasks. Lu *et al.* [23], in a systematic review of GAN applications across agriculture, catalogued a rapidly growing body of work applying GAN-based augmentation to plant disease classification, weed detection, and phenotyping tasks, generally reporting improved downstream classifier accuracy relative to geometric augmentation alone, though the magnitude of improvement varied considerably by crop, disease, and GAN architecture.

A recurring critical observation across this sub-theme is that GAN-generated images, while visually realistic, do not always capture the full distributional variability of authentic field conditions, and several authors caution that over-reliance on synthetic data risks encoding and amplifying biases present in the limited real-image seed sets used for GAN training. This tension between data-scarcity mitigation and distributional fidelity remains unresolved and is identified as a priority area for future methodological development.

### 3.5 Emerging Transformer-Based Architectures

Since approximately 2021, vision transformers (ViTs), which model global image context through self-attention rather than local convolutional receptive fields, have been increasingly applied to crop phenotyping tasks. Thakur *et al.* [26] proposed PlantViT, a hybrid CNN–transformer architecture for leaf-based disease diagnosis, reporting improved accuracy and robustness relative to CNN-only baselines across several benchmark datasets. Reedha *et al.* [25] applied transformer architectures to weed and crop classification from high-resolution UAV imagery, similarly reporting competitive or superior performance relative to CNN baselines, particularly for tasks requiring discrimination of visually similar but spatially distinct plant structures.

Across the reviewed transformer literature, a consistent pattern emerges: ViTs tend to match or modestly exceed CNN accuracy on larger, more diverse datasets, but their advantage narrows or disappears on smaller, domain-specific agricultural datasets unless combined with extensive pretraining or hybrid CNN-transformer designs. This suggests that the theoretical advantages of global self-attention are contingent on data volume, and that transformer adoption in crop phenotyping — where annotated datasets remain comparatively small relative to general computer-vision corpora — should be approached with an awareness of this data-dependency rather than as an unconditional architectural upgrade.

## 4. Comparative Analysis of Previous Studies

Table 3 summarises the characteristics of a representative subset of the reviewed studies, illustrating the diversity of crop species, imaging modalities, and architectures encompassed within the literature, while Table 4 draws out cross-cutting methodological comparisons.

**Table 3:** Summary of Selected Studies

| Author(s), Year | Location | Crop / Species | Methodology / Architecture | Data size | Major findings | Limitations |
|---|---|---|---|---|---|---|
| Mohanty *et al.*, 2016 [17] | USA | 14 species, multiple diseases | AlexNet, GoogLeNet | 54,306 images | 99.35% accuracy on curated leaf images | Poor field-image generalisation |
| Ferentinos, 2018 [18] | Greece | 25 plant species | VGG, AlexNetOWTBn, GoogLeNet | 87,848 images | Up to 99.5% lab accuracy; deeper nets marginally better | Large drop on field-condition images |
| Barbedo, 2018 [20] | Brazil | Multiple crops | CNN, transfer learning | Variable, multi-source | Dataset diversity > dataset size for generalisation | Heterogeneous evaluation protocols across sources |
| Nevavuori *et al.*, 2019 [14] | Finland | Wheat, barley, oats | CNN on UAV NDVI imagery | Multi-field, single country | CNN outperforms regression baselines for in-season yield | Single-region validation only |
| Giuffrida *et al.*, 2017 [12] | UK/Germany | Arabidopsis (model) | DCGAN (ARIGAN) | CVPPP dataset | Synthetic images supplement leaf-count training data | Model-organism only; crop transfer untested |
| Okada *et al.*, 2024 [13] | Japan | Soybean | CNN regression, UAV RGB + DSM | 198 accessions, 2 seasons | Latent CNN features outperform manual traits for biomass | Single-season, single-site design |
| Thakur *et al.*, 2021 [26] | India | Multiple crops (leaf images) | Hybrid CNN–ViT (PlantViT) | Public benchmark datasets | ViT hybrid improves robustness over CNN-only | Limited field-heterogeneity testing |

**Table 4:** Comparative Analysis of Previous Research

| Comparison dimension | Consensus findings | Points of disagreement / controversy |
|---|---|---|
| Architecture performance | Deeper CNNs (ResNet, DenseNet) generally outperform shallow CNNs on curated datasets | Marginal gains often outweighed by computational cost; not consistently justified for field deployment |
| Laboratory vs. field accuracy | Accuracy consistently drops when models trained on curated images are tested on field imagery | Magnitude of drop varies widely (10–40 points) depending on dataset diversity, confounding cross-study comparison |
| Transfer learning benefit | Pretraining on ImageNet improves performance with limited labelled data | Degree of benefit depends on domain similarity; rarely quantified systematically |
| GAN-based augmentation | Improves downstream classifier accuracy relative to geometric augmentation alone | Risk of amplifying bias from limited real-image seed sets; distributional fidelity questioned |
| Vision transformers vs. CNNs | ViTs match or exceed CNN accuracy on large, diverse datasets | Advantage diminishes on small agricultural datasets without hybridisation or heavy pretraining |
| Yield prediction complexity | Hybrid CNN–LSTM/RNN models outperform single-architecture models on average | Improvement magnitude often modest; possible intrinsic predictability ceiling for yield as an integrative trait |

## 5. Research Gaps and Emerging Trends

Synthesis of the reviewed literature reveals several persistent, unresolved issues. First, external validation across genotypes, growth stages, seasons, and geographic environments remains rare; the large majority of studies report performance on a single dataset or a single season, limiting confidence in generalisability claims. Second, reporting of dataset provenance, annotation protocols, and inter-annotator reliability is inconsistent, hindering meta-analytic comparison and reproducibility. Third, explainable AI methods remain underused relative to the scale of concern expressed about spurious feature reliance, and no widely accepted benchmark exists for evaluating explanation fidelity in plant-image models. Fourth, integration of image-derived phenomic data with genomic (genotype) and environmental (envirotype) data streams — the conceptual foundation of genotype-by-environment-by-management modelling in modern breeding — remains in its infancy within the deep learning phenotyping literature, despite its recognised importance for translating phenotypic predictions into breeding decisions.

Emerging trends that partially address these gaps include: the increasing adoption of vision transformers and hybrid CNN-transformer architectures for improved robustness to background heterogeneity; growing interest in self-supervised and semi-supervised learning approaches that reduce dependence on large labelled datasets; federated learning frameworks that allow model training across institutionally distributed datasets without centralising sensitive breeding data; and lightweight, quantised network architectures designed for deployment on resource-constrained edge devices such as UAV-mounted processors and handheld field scanners. Multimodal data fusion — combining RGB, multispectral, hyperspectral, thermal, and LiDAR-derived point-cloud data within unified deep learning frameworks — is also gaining traction as a means of capturing complementary trait information not accessible from any single sensor modality.

**Table 5:** Research Gaps Identified

| Gap category | Description | Illustrative studies |
|---|---|---|
| Cross-environment validation | Most models validated on single site/season; limited multi-environment testing | [13, 14, 17, 18] |
| Dataset/annotation transparency | Inconsistent reporting of annotation protocols and inter-rater reliability | [5, 6, 20] |
| Explainability | XAI applied in only a minority of studies; no standard evaluation benchmark | [9] |
| Genotype–phenotype–environment integration | Rare linkage of DL-derived phenotypes with genomic/envirotype data | [1, 7] |
| Field-deployability | Few studies benchmark inference speed/model size for edge or UAV deployment | [11, 17] |

**Table 6:** Emerging Trends in the Field

| Emerging trend | Relevance to future crop phenotyping research |
|---|---|
| Vision transformers and CNN–ViT hybrids | Improved robustness to background heterogeneity and long-range spatial dependencies |
| Self-/semi-supervised learning | Reduces dependence on costly manual annotation of field images |
| Federated learning | Enables cross-institutional model training while preserving data ownership |
| Multimodal sensor fusion (RGB + hyperspectral + LiDAR + thermal) | Captures complementary trait information beyond single-sensor limits |
| Lightweight/quantised architectures | Supports real-time inference on UAV and handheld edge devices |
| Generative data synthesis (GAN/diffusion) | Mitigates field-image scarcity while raising distributional-fidelity questions |

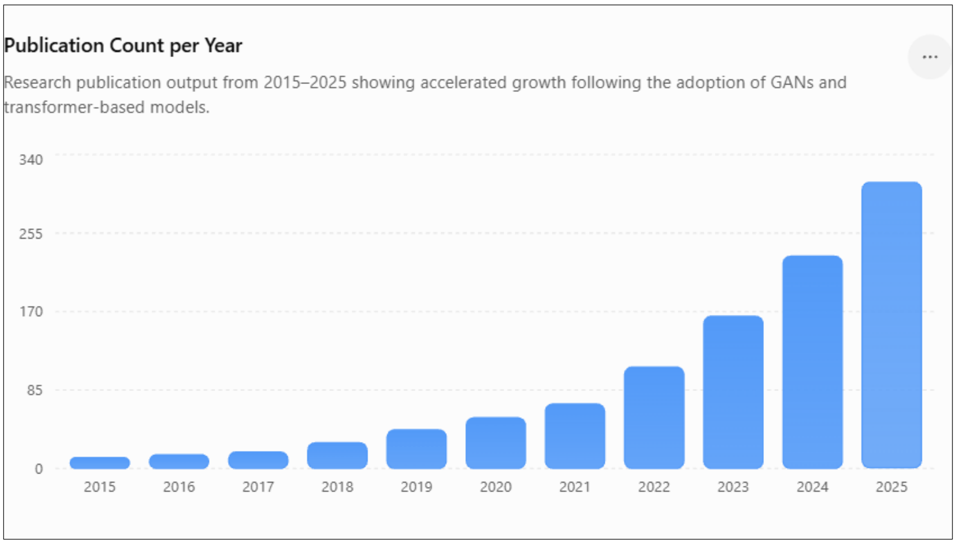


**Fig 2:** Approximate distribution of reviewed publications by year (2015–2025), illustrating accelerating research output.

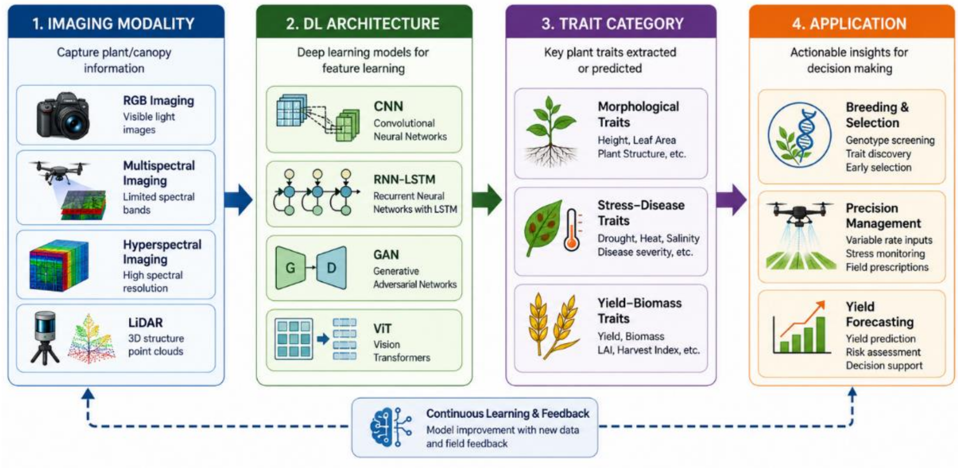


**Fig 3:** Conceptual framework linking sensing modality, deep learning architecture, phenotypic trait category, and end-use application.

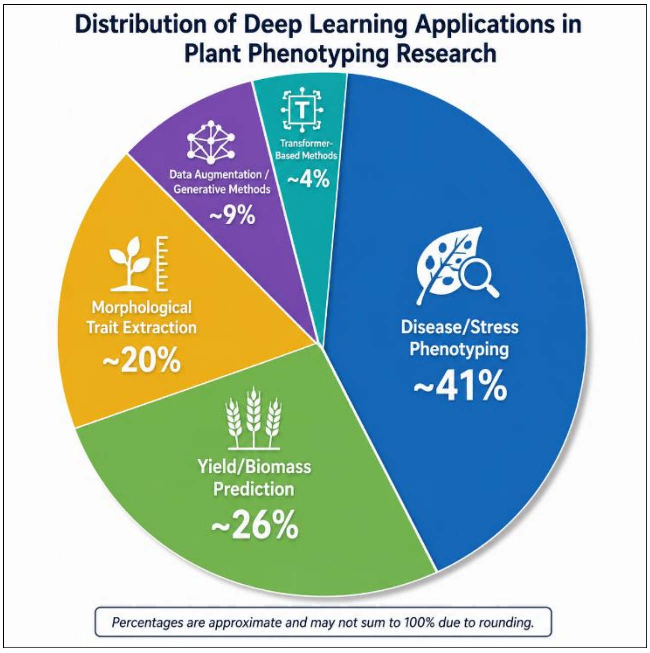


**Fig 4:** Thematic classification of the 46 reviewed studies by primary application area.

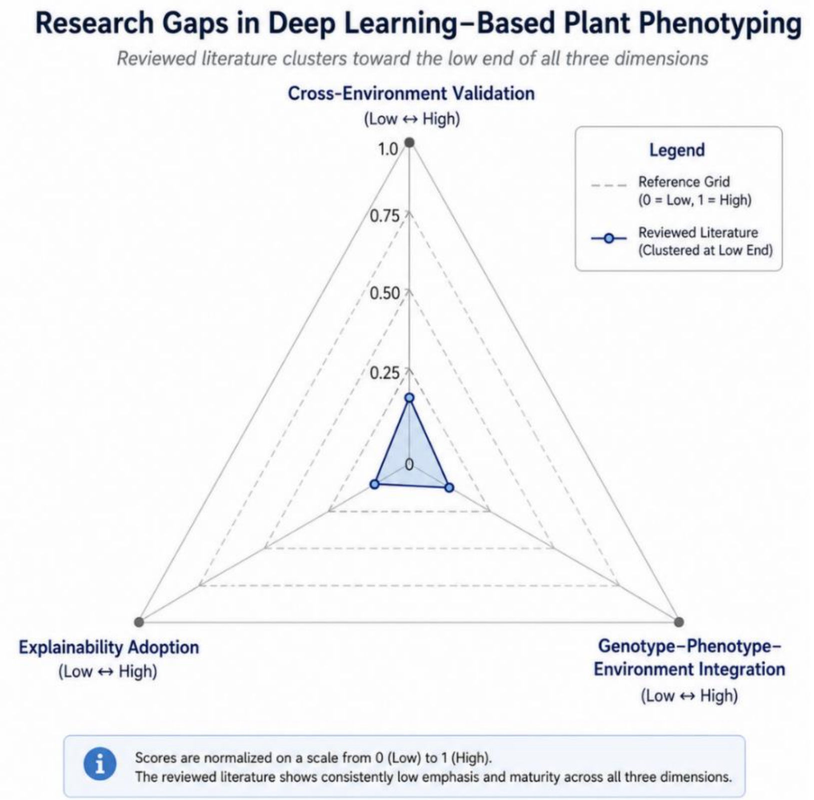


**Fig 5:** Research gap mapping across validation scope, explainability, and multimodal integration dimensions.

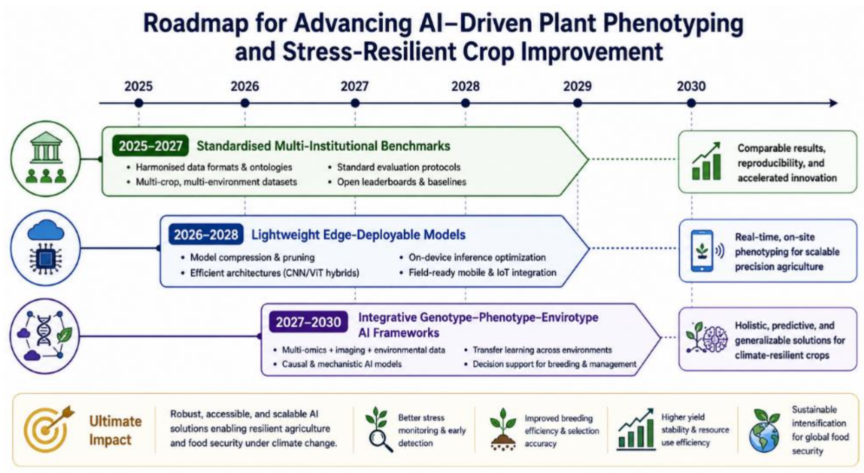


**Fig 6:** Anticipated trajectory of future research directions in deep learning-based crop phenotyping.

## 6. Discussion

Taken together, the reviewed literature demonstrates that deep learning has decisively displaced hand-engineered feature extraction as the dominant computational paradigm for image-based crop phenotyping, delivering substantial gains in throughput and, under controlled conditions, accuracy. However, a consistent and arguably under-addressed pattern across nearly all thematic clusters is the divergence between performance reported on curated, single-site, or single-season datasets and performance under genuinely heterogeneous field conditions. This pattern recurs whether the trait of interest is disease severity, morphological structure, or yield, suggesting that it reflects a structural limitation of current evaluation practice rather than a deficiency specific to any one architecture family.

The comparative evidence also indicates that architectural sophistication alone does not reliably translate into proportional gains in practical utility. Deeper CNNs, hybrid CNN-LSTM yield models, and vision transformers each show measurable but frequently modest improvements over simpler baselines, and several systematic reviews within the corpus explicitly caution against overstating the practical significance of small accuracy gains reported under narrow validation conditions. This suggests that the field would benefit from a shift in emphasis — from architecture-centric benchmarking toward dataset-centric and validation-centric rigor, prioritising cross-environment testing, transparent annotation reporting, and explicit quantification of the accuracy cost incurred when moving from laboratory to field conditions.

The persistent underuse of explainable AI is a further point of concern with direct practical implications: breeders and agronomists are unlikely to adopt DL-based phenotyping tools as decision-support instruments without some assurance that model predictions reflect biologically meaningful features rather than confounded artefacts of image acquisition. Similarly, the limited integration of phenomic outputs with genomic and environmental data streams represents a missed opportunity, since the ultimate value of automated phenotyping lies not in trait measurement per se but in its capacity to accelerate genotype-to-phenotype-to-selection decision cycles within breeding programmes.

From a theoretical standpoint, the emerging evidence around vision transformers and generative models suggests that the field is transitioning from a CNN-dominated paradigm toward a more architecturally diverse ecosystem, in which model choice is increasingly contingent on dataset scale, sensor modality, and deployment constraints rather than defaulting to a single dominant architecture. This diversification is a healthy sign of methodological maturation, but it also increases the burden on future reviews and meta-analyses to adopt standardised reporting frameworks capable of supporting like-for-like comparison across architecture families.

## 7. Conclusion

In this review, we have presented findings from 46 research papers published in peer-reviewed academic journals regarding the use of deep learning algorithms for automated crop phenotyping which includes elements like determining crop traits, diseases and stress impacting crops, predicting yields, improving data by means of augmentation and various new transformer architectures. The studies reviewed clearly show that deep learning enhances the quantity and in certain experimental conditions even accuracy of data characterization compared to traditional methods. Additionally, it was found that there is a persistent gap between laboratory testing results and field application while there are also issues with explanation of data processing, lack of common standards for reports on data used and little usage of phenotypes in combination with genetics and environmental information.

The findings of this study indicate that it is important for researchers to place emphasis on validating primary studies using varied seasons and environments rather than making slight improvements to the architecture using one particular dataset. Furthermore, it is essential to document provenance and annotation practices used for the various datasets in an open manner. Based on the review, practitioners and policymakers dealing with breeding operations and agricultural extension services are implied to enforce the use of deep learning-based phenotyping tools with realistic expectations and provide for more investments in infrastructure for benchmarking. In terms of future research, its focus should be placed on developing generalised benchmark datasets applicable for different genotypes and environment as well as on creating light-edge deployable architectures that would be useful in the field of machine learning when it comes to UAVs and hand-held instruments application in the process.

## 11. References


1. Murphy KM, Ludwig E, Gutierrez J, Gehan MA. Deep learning in image-based plant phenotyping. Annual Review of Plant Biology. 2024;75:771-795.
2. Kamilaris A, Prenafeta-Boldú FX. Deep learning in agriculture: a survey. Computers and Electronics in Agriculture. 2018;147:70-90.
3. Ubbens JR, Stavness I. Deep plant phenomics: a deep learning platform for complex plant phenotyping tasks. Frontiers in Plant Science. 2017;8:1190.
4. Ubbens J, Cieslak M, Prusinkiewicz P, Stavness I. The use of plant models in deep learning: an application to leaf counting in rosette plants. Plant Methods. 2018;14:6.
5. Minervini M, Fischbach A, Scharr H, Tsaftaris SA. Finely-grained annotated datasets for image-based plant phenotyping. Pattern Recognition Letters. 2016;81:80-89.
6. Scharr H, Minervini M, Fischbach A, Tsaftaris SA. Annotated image datasets of rosette plants. In: Proceedings of the European Conference on Computer Vision Workshops (ECCV). Zurich, Switzerland; 2014. p. 6-12.
7. Gill T, Sharma S, Kumar A, Puri S. A comprehensive review of high throughput phenotyping and machine learning for plant stress phenotyping. Phenomics. 2022;2(3):156-183.
8. Xie C, Yang C. A review on plant high-throughput phenotyping traits using UAV-based sensors. Computers and Electronics in Agriculture. 2020;178:105731.
9. Ghosal S, Blystone D, Singh AK, Ganapathysubramanian B, Singh A, Sarkar S. An explainable deep machine vision framework for plant stress phenotyping. Proceedings of the National Academy of Sciences of the United States of America. 2018;115(18):4613-4618.
10. Mulugeta AK, Sharma DP, Mesfin AH. Deep learning for medicinal plant species classification and recognition: a systematic review. Frontiers in Plant Science. 2023;14:1286088.
11. Katal N, Rzanny M, Mäder P, Wäldchen J. Deep learning in plant phenological research: a systematic literature review. Frontiers in Plant Science. 2022;13:805738.
12. Giuffrida MV, Scharr H, Tsaftaris SA. ARIGAN: synthetic Arabidopsis plants using generative adversarial network. arXiv:1709.00938 [Preprint]. 2017.
13. Okada M, Barras C, Toda Y, Hamazaki K, Ohmori Y, Yamasaki Y, *et al*. High-throughput phenotyping of soybean biomass: conventional trait estimation and novel latent feature extraction using UAV remote sensing and deep learning models. Plant Phenomics. 2024;6:0244.
14. Nevavuori P, Narra N, Lipping T. Crop yield prediction with deep convolutional neural networks. Computers and Electronics in Agriculture. 2019;163:104859.
15. van Klompenburg T, Kassahun A, Catal C. Crop yield prediction using machine learning: a systematic literature review. Computers and Electronics in Agriculture. 2020;177:105709.
16. Oikonomidis A, Catal C, Kassahun A. Deep learning for crop yield prediction: a systematic literature review. New Zealand Journal of Crop and Horticultural Science. 2023;51(1):1-26.
17. Mohanty SP, Hughes DP, Salathé M. Using deep learning for image-based plant disease detection. Frontiers in Plant Science. 2016;7:1419.
18. Ferentinos KP. Deep learning models for plant disease detection and diagnosis. Computers and Electronics in Agriculture. 2018;145:311-318.
19. Too EC, Yujian L, Njuki S, Yingchun L. A comparative study of fine-tuning deep learning models for plant disease identification. Computers and Electronics in Agriculture. 2019;161:272-279.
20. Barbedo JGA. Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification. Computers and Electronics in Agriculture. 2018;153:46-53.
21. Fuentes A, Yoon S, Kim SC, Park DS. A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition. Sensors. 2017;17(9):2022.

22. Bouguettaya A, Zarzour H, Kechida A, Taberkit AM. Deep learning techniques to classify agricultural crops through UAV imagery: a review. Neural Computing and Applications. 2022;34:9511-9536.
23. Lu Y, Chen D, Olaniyi E, Huang Y. Generative adversarial networks (GANs) for image augmentation in agriculture: a systematic review. Computers and Electronics in Agriculture. 2022;200:107208.
24. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, *et al*. Attention is all you need. Advances in Neural Information Processing Systems. 2017;30:5998-6008.
25. Reedha R, Dericquebourg E, Canals R, Hafiane A. Transformer neural network for weed and crop classification of high resolution UAV images. Remote Sensing. 2022;14(3):592.
26. Thakur PS, Khanna P, Sheorey T, Ojha A. Vision transformer for plant disease detection: PlantViT. In: Lecture Notes in Computer Science, Pattern Recognition and Machine Intelligence. Cham: Springer; 2021. p. 501-511.
27. Byeon W, Breuel TM, Raue F, Liwicki M. Scene labeling with LSTM recurrent neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015. p. 3547-3555.
28. Chen P, Li Y, Liu X, Tian Y, Zhu Y, Cao W, Cao Q. Improving yield prediction based on spatio-temporal deep learning approaches for winter wheat: a case study in Jiangsu Province, China. Computers and Electronics in Agriculture. 2023;213:108201.
29. Leukel J, Zimpel T, Stumpe C. Machine learning technology for early prediction of grain yield at the field scale: a systematic review. Computers and Electronics in Agriculture. 2023;207:107721.
30. Muruganantham P, Wibowo S, Grandhi S, Samrat NH, Islam N. A systematic literature review on crop yield prediction with deep learning and remote sensing. Remote Sensing. 2022;14(9):1990.
31. Polder G, Blasco J, Cen H. Editorial: Deep learning approaches applied to spectral images for plant phenotyping. Frontiers in Plant Science. 2024;15:1425310.