Enablers of Agriculture 4.0: A Comprehensive Survey of Machine Learning, Robotics, and Emerging Technologies Driving the Next Agricultural Paradigm

Kiran K. Kethineni [1,*], Rishi Raj Kanukuntla [1], Saraju P. Mohanty [1], Elias Kougianos [2]

[1] Department of Computer Science and Engineering, University of North Texas, Denton, TX, USA
[2] Department of Electrical Engineering, University of North Texas, Denton, TX, USA

{rishirajkanukuntla@my.unt.edu, saraju.mohanty@unt.edu, elias.kougianos@unt.edu}
* Correspondence: kirankumar.kethineni@unt.edu

Abstract

Agriculture 4.0 is usually presented through its most visible products: autonomous tractors, weed-spotting drones, and language-driven advisory apps. Field deployment has proven harder than these demonstrations suggest. Models that report accuracies above 99 % on curated leaf images frequently lose 25–40 percentage points when re-run on smartphone photographs from a working farm; reinforcement-learning irrigation controllers that save 30 % water in simulation routinely over- or under-water clay and sand soils that violate the simulator's homogeneity assumptions; and the machine-learning stack actually deployed across hundreds of millions of hectares remains dominated by gradient-boosted trees rather than the transformer architectures dominating the literature. This survey is organised around this gap between reported and deployed performance. We review the four agricultural revolutions, describe the hardware, connectivity, data, and intelligence layers of the modern stack, and map eleven model families onto eight application domains that cover most field-level decisions. Two analytical sections address deployment directly: one compares model families on the criteria that govern model choice in practice (latency, edge-feasibility, data efficiency, interpretability), and one documents the lab-to-field failure modes that benchmark scores hide. We close with seven research frontiers (geometric deep learning, agricultural foundation models, digital twins, agentic systems, continual learning, edge AI, federated data) and seven deployment barriers (domain shift, calibration, interpretability, smallholder cost, regulation, sovereignty, robustness). The survey is aimed at two readers: researchers entering the area who want a reliable map of what works, and practitioners who need to know where reported numbers should be treated with caution.

Keywords: Agriculture 4.0; precision agriculture; deep learning; edge AI; vision transformers; large language models; federated learning; foundation models; robotic perception; lab-to-field gap

1 Introduction

By 2050, food production will need to support roughly two billion more people on a land base that is no longer expanding, with a shrinking agricultural labour force across much of the world and a climate whose recent history is a poor guide to the coming decade [1]. The conventional levers of productivity, larger machines, more synthetic inputs, and improved genetics, have been used heavily for sixty years, and the largest remaining gains are informational: applying the right inputs to the right areas at the right times. This is the central idea of Agriculture 4.0 [2]. It shifts the guiding question from how to apply more inputs toward how to sense conditions, decide, and act with the accuracy of a trained agronomist, across a whole farm and throughout the season.

The technology needed to answer that question has matured over the past decade. Dense in-field sensors and aerial multispectral imagery now provide data at resolutions and revisit rates that were not available in the precision-agriculture era [3]; low-power edge accelerators run convolutional and transformer inference within a 15 W budget; instruction-tuned language models provide a natural-language interface between farmers and decision systems; and federated learning offers a way to improve models across farms without requiring growers to share management data [4]. None of these capabilities was available to a precision-agriculture deployment in 2010.

What has not matured is confidence that a given published method will work on an actual farm. The literature contains many high-accuracy benchmark results that drop sharply, sometimes by 30 percentage points or more, when the same models are re-evaluated on held-out sites, held-out years, or smartphone-quality images [5, 6]. The reasons are familiar: distribution shift across regions and seasons, class imbalance for rare but economically important pests and diseases, evaluation of robot policies only in simulation, and overconfident model outputs driving actions such as chemical application, irrigation, and harvesting, where mistakes are costly or irreversible. A survey of this area therefore needs to do two things: report the methods in the literature, and indicate which of them the deployment record does not yet support.

1.1 Scope and Contributions

We focus on the machine-learning, robotics, and sensing technologies that play an active role in the perception–reasoning–action loop of Agriculture 4.0. This includes classical machine learning, deep learning across modalities, large language models, and the more recent foundation-model and federated-learning work. We do not cover purely statistical agronomic modelling, controlled-environment indoor farming, or post-harvest logistics, except where they illustrate a point that transfers. The survey makes five main contributions:

1. **A consolidated comparison with the seven prior surveys most cited on AI for agriculture**, summarised in Table 1. The table shows that no prior survey covers classical ML, deep learning, large language models, robotics, IoT/edge, and forward-looking research directions in a single source, and no prior survey examines deployment realism as a structural concern.

2. **A model-family capability matrix** (Table 2) that scores eleven architectural families on the dimensions that actually drive selection in deployment: real-time feasibility, edge-feasibility, few-shot ability, native interpretability, spectral support, temporal support, and the application areas where the family carries its weight.

3. **A curated index of ten benchmark datasets** (Table 3) covering disease, weed, pest, fruit, soil, satellite, and advisory tasks, with the modality, scale, and representative models that have used each.

4. **A per-domain method ledger** (Table 4) that records, for each of eight application domains, the canonical references, the model architecture used, the dataset evaluated against, and the reported metric, structured so that benchmark-only entries are easy to identify.

5. **Two analytical sections** that are uncommon in comparable surveys: §4.10 examines when each model family actually delivers and where its reported accuracy is systematically inflated, and §5.9 catalogues the lab-to-field failure modes that the prevailing evaluation protocols conceal.

Table 1. Comparison of this survey with representative prior AI/ML-in-agriculture surveys.

Survey	Year	Class. ML	Deep Learning	LLMs/FM	Robotics	IoT/Edge	Future Dir.	Domains
Liakos et al. [7]	2018	✓	○	—	—	○	—	4
Kamilaris & Prenafeta-Boldú [8]	2018	—	✓	—	—	—	—	6
Patrício & Rieder [9]	2018	○	✓	—	○	—	—	5
Koirala et al. [10]	2019	—	✓	—	○	—	—	2
Saleem et al. [11]	2019	○	✓	—	—	—	—	2
Fountas et al. [12]	2020	—	○	—	✓	○	○	4
Benos et al. [13]	2021	✓	✓	—	—	○	○	5
This survey	**2026**	✓	✓	✓	✓	✓	✓	**8**

✓ = covered in depth; ○ = partially covered; — = not covered.

2 Historical Arc of Agricultural Automation

Over roughly a hundred and fifty years, agriculture has passed through four distinct revolutions. The label Agriculture 4.0 is useful because each earlier transition removed a different bottleneck, and the current one is the first to address human judgement rather than human labour or agricultural chemistry. Figure 1 summarises the sequence.

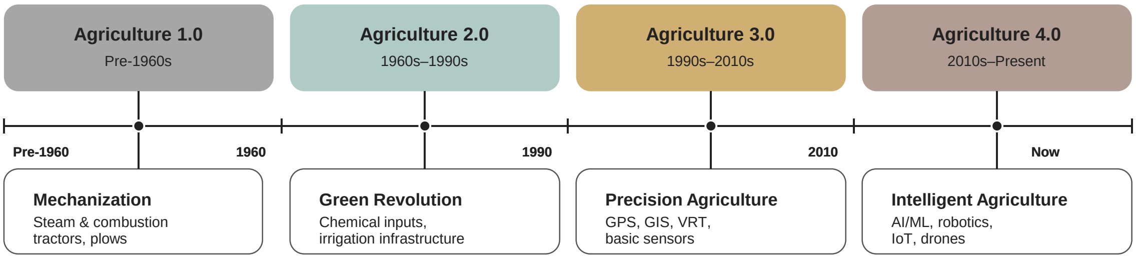

Figure 1. The four agricultural revolutions.

2.1 Agriculture 1.0: Mechanisation (Pre-1960s)

The first revolution replaced draft animals and manual labour with machinery. Steam ploughs in the 1850s, the gasoline tractor in the 1900s [14], and the self-propelled combine in the 1930s each removed a labour bottleneck rather than an information one: the machines did more work per hour, but every agronomic decision still rested with the operator. Fields were managed as uniform units, and spatial variation finer than "which corner of the farm" was not part of management.

2.2 Agriculture 2.0: The Green Revolution (1960s–1990s)

The second revolution came from biology and chemistry rather than computing. High-yielding semi-dwarf wheat and rice varieties, large-scale ammonia fertilisation, broad-spectrum herbicides, and reservoir-fed irrigation raised cereal output substantially, particularly across South Asia and Latin America [15]. Information was handled on paper: extension agents gave advice that was uniform across whole districts, and feedback to research stations arrived as season-end yield totals. The environmental costs of uniformly high inputs, including salinisation, nitrate pollution, biodiversity loss, and herbicide resistance, accumulated over this period and created much of the pressure that the next revolution had to address.

2.3 Agriculture 3.0: Precision Agriculture (1990s–2010s)

The third revolution introduced the idea that a field is not uniform. Civilian GPS in the early 1990s linked farm machinery to absolute coordinates [16]; Variable Rate Technology allowed fertiliser, seeding, and pesticide rates to vary within a single pass; Geographic Information Systems provided a coordinate system for the data; and satellite vegetation indices (NDVI, EVI from Landsat and MODIS) made biomass differences visible at landscape scale. What this era did not do was close the loop. The workflow was sense, map, print a prescription, have a person read it, and have a person act. Automation stopped at the prescription map, and decision-making stayed with the operator.

2.4 Agriculture 4.0: Intelligent and Autonomous Agriculture (2010s–Present)

Agriculture 4.0 is the era in which the loop closes. Five developments arrived within a few years of each other, so that by the mid-2010s a working prototype could perceive, reason, and act at field speed without continuous human supervision:

- Affordable multi-rotor UAVs after roughly 2012 brought a flexible aerial sensing platform to fields of any size [17].
- AlexNet's 2012 ImageNet result [18] began a decade of deep-learning progress that transferred readily to plant phenotyping and agricultural perception.
- Embedded GPU-class accelerators (NVIDIA Jetson family, later Hailo, Coral, RB5) made real-time neural inference viable on a tractor or a quadrotor.
- Low-power wide-area protocols (LoRaWAN, NB-IoT) made dense field sensor networks deployable without per-node cellular subscriptions or power infrastructure [19].
- Instruction-tuned language models, post-2022, gave the user-facing layer of the stack a way to talk to non-expert farmers in their own language.

The defining feature of the present era is not any single technology but the closure of the perception–reasoning–action loop, with humans supervising rather than operating. The rest of this survey examines that loop in detail.

3 The Agriculture 4.0 Technology Stack

Describing Agriculture 4.0 as a "stack" reflects real dependencies between its layers. The available sensing modality constrains which models can run; the available connectivity constrains where those models can run; and the data layer determines whether they can be trained at all. Figure 2 shows the arrangement used in the rest of the paper. The lower brace marks the parts of the system that interact with the physical world through sensors and actuators, and the upper brace marks the decision and interface layers that use their output.

3.1 Hardware Layer

The hardware layer is where many failures originate, and where many methods that work in principle are ruled out by power or weight limits.

Aerial platforms. Multi-rotor UAVs with payloads of 1–5 kg now routinely carry RGB, five-to-ten-band multispectral, hyperspectral (200+ bands), thermal, and compact LiDAR sensors [3]. Multispectral imagery is the most widely used: vegetation indices (NDVI, NDRE, SAVI) correlate well enough with nitrogen status, water deficit, and several disease proxies to be operationally useful, despite their saturation at high biomass. Hyperspectral imagery provides more information

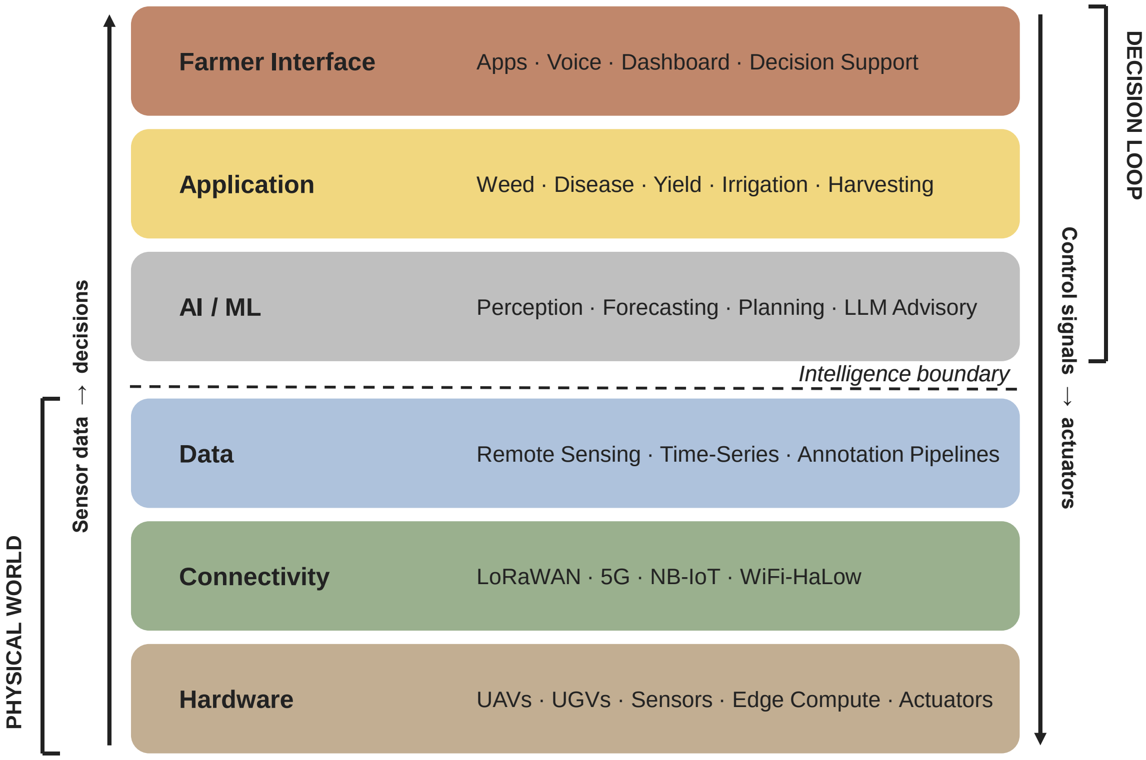

Figure 2. The Agriculture 4.0 technology stack.

but needs a larger payload and much more difficult calibration, and it remains research-grade in most deployments. LiDAR is increasingly used for 3D canopy reconstruction and biomass estimation, and falling sensor cost is making it viable on sub-kilogram drones.

Ground platforms. UGVs perform the tasks that require contact with plants: mechanical weeding, banded spraying, soil sampling, and selective harvesting [12, 20]. The most common recent design is UAV/UGV cooperation, in which aerial imaging builds a geo-referenced task map that a ground rover then executes. The Flourish system [21] is a well-developed example: an aerial multispectral survey produces a geo-referenced weed map that directs a ground rover for selective spraying and mechanical weeding. The reduction in herbicide use comes from targeting treatment at the patch level rather than from any particular improvement in the detector.

In-field IoT nodes. The standard sensor set includes soil-moisture probes at several depths, soil EC, leaf wetness, ambient temperature and humidity, solar radiation, and wind speed. The main design constraint is power. A node that needs a battery change every season is a maintenance burden across a deployment of several thousand nodes, so energy-harvesting nodes (small solar or thermal-gradient) and aggressive duty-cycling are common. This constraint carries up into the ML layer as a preference for models that tolerate missing or stale features.

Edge accelerators. Closed-loop perception for weed spraying, harvesting, and obstacle avoidance cannot tolerate the latency of a cloud round-trip. The current generation of accelerators is dominated by NVIDIA Jetson AGX/Orin (15–60 W), Hailo-8/8M (2.5 W at 26 TOPS), Qualcomm RB5, and Google Coral TPU. The binding constraint is rarely raw TOPS; more often it is thermal throttling under direct sun, which is seldom tested in published work.

3.2 Connectivity Layer

LoRaWAN handles the long-range (>5 km), low-bandwidth telemetry that sensor networks require [19]. Where cellular coverage exists, 5G and NB-IoT carry high-bandwidth robot video, over-the-air model updates, and teleoperation. WiFi-HaLow (802.11ah) occupies a less-discussed middle ground: medium range, moderate bandwidth, and a good fit for mesh networks between sensor nodes and a farm gateway. The absence of a single dominant agricultural networking standard adds real engineering cost, and surveys that omit the connectivity layer tend to underestimate its effect on system reliability.

3.3 Data Layer

Remote sensing baselines. Sentinel-2 (10 m, 5-day revisit) is the default free imagery; Planet SkySat (sub-metre, daily) and Maxar (very-high-resolution, on request) cover cases where it is insufficient; and MODIS (500 m, daily) is still widely used for regional yield modelling. None of these resolutions matches a UAV overflight, but their revisit frequency is one that UAVs cannot match in practice.

The annotation bottleneck. Agricultural computer vision is harder than ImageNet-style computer vision in ways that do not appear in a FLOP count. Labelled data is scarce, expensive, and rarely representative. PlantVillage [22] contains tens of thousands of images, but most are leaves on a uniform background under controlled lighting. Models that reach 99 % or more on that data generalise poorly to leaves in a dusty smartphone photograph taken in the field [5, 6]. Much of the reported progress on agricultural classification benchmarks over the past decade reflects improvement under these controlled conditions rather than improvement in field performance.

3.4 AI and ML Layer

We divide the discussion of this layer across the next two sections to keep model families separate from their uses. Section 4 covers the model families. Section 5 covers what those models are used for. The closing subsection of each (§4.10 and §5.9) assesses what the families and applications actually deliver in practice.

4 Machine Learning Models in Agriculture 4.0

The agricultural literature contains a very large number of model architectures, but in deployment they are best organised by the role each family plays in the closed loop. We use the Perception–Reasoning–Action (PRA) decomposition in Figure 3. Perception is dominated by CNNs and ViTs; reasoning combines classical regressors, recurrent and state-space models, GNNs, and LLMs; and action is dominated by RL policies and motion planners. The dashed feedback arrow marks the step that the precision-agriculture era never closed.

Within each PRA stage, model selection depends less on raw accuracy than on a few cross-cutting properties: inference latency, edge feasibility, data efficiency, and interpretability. Figure 4 groups the eleven model families found in the agricultural literature into five branches, from classical ML to foundation models, to make these trade-offs explicit.

Table 2 maps these five branches onto seven deployment-relevant dimensions, so that model selection can be based on operational constraints rather than benchmark rankings alone.

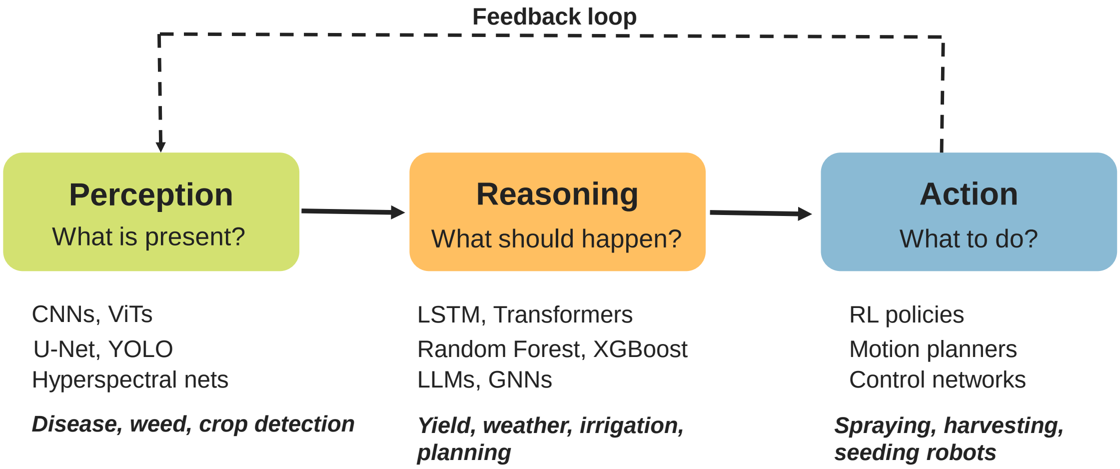

Figure 3. The Perception–Reasoning–Action (PRA) loop.

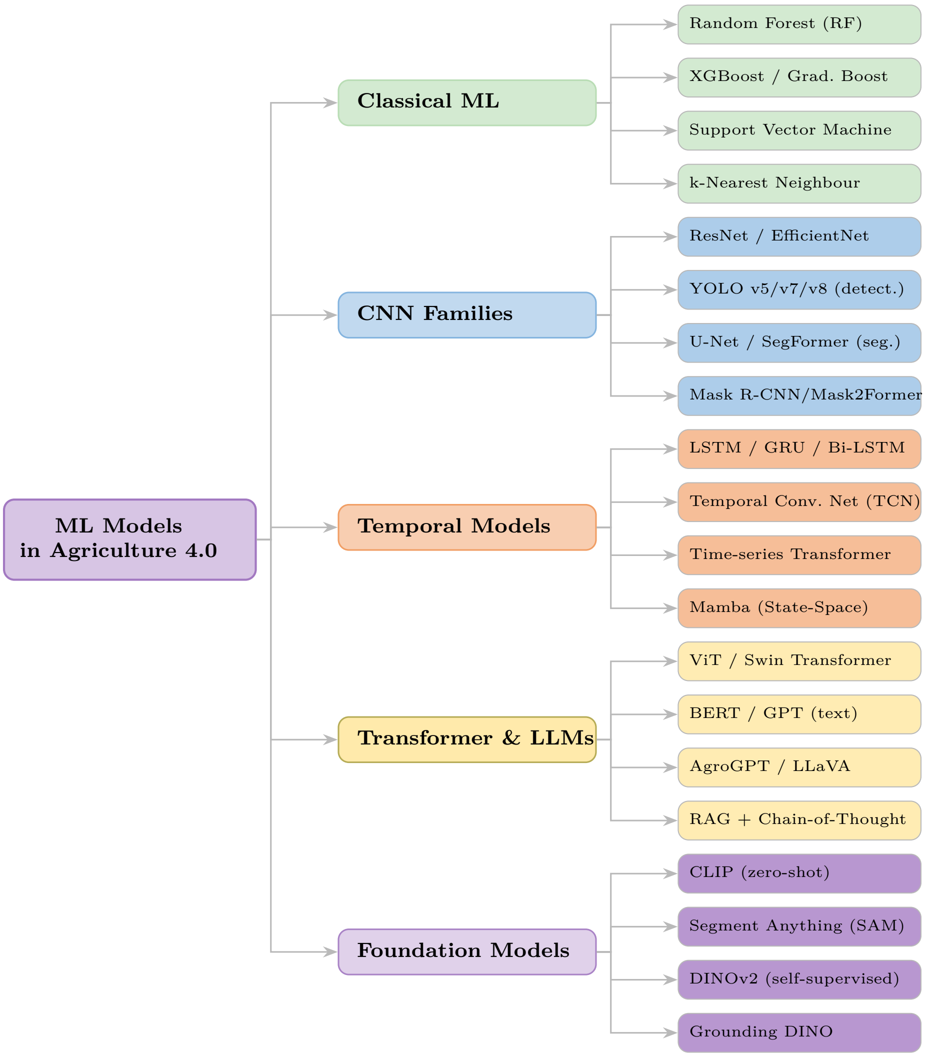

Figure 4. Taxonomy of the five model families.

Table 2. Model-family capability matrix across seven deployment dimensions.

Model Family	**Real-time**	**Edge**	**Few-shot**	**Interpretable**	**Spectral**	**Temporal**	**Where it carries its weight**
Classical ML (RF, SVM, XGB)	✓	✓	○	✓	✓	—	Yield, soil, crop mapping
CNN — Classification	✓	○	○	○	✓	—	Disease, species ID
CNN — Detection (YOLO)	✓	○	—	○	✓	—	Weed, pest, fruit detection
CNN — Segmentation (U-Net)	○	○	—	○	✓	—	Canopy, field mapping
LSTM / GRU / Mamba	○	○	—	—	○	✓	Yield, weather, soil moisture
ViT / Swin Transformer	○	—	✓	○	✓	○	Fine-grained classification
GNN (spatial-temporal)	○	—	○	○	○	✓	Field-network propagation
RL (DQN, PPO, MARL)	✓	○	—	—	—	✓	Irrigation, harvesting
LLMs (AgroGPT, GPT-4)	—	—	✓	✓	—	○	Advisory, planning, RAG
Foundation Models (CLIP, SAM)	○	—	✓	○	✓	—	Open-vocabulary detection, segmentation
Diffusion / GAN	—	—	✓	—	✓	—	Data augmentation, synthesis

✓ = yes; ○ = partial or lite version; — = no.

4.1 Classical Machine Learning

Classical methods such as random forests, gradient boosting, and support vector machines are not merely a preliminary to deep learning. On the inputs that an Agriculture 4.0 stack produces in volume (multi-band spectral readings, soil and weather sensor streams, and satellite-derived indices), they remain competitive or better, and they offer properties that neural networks still struggle to match: low training cost, calibrated probability outputs, built-in handling of missing features, and interpretable attributions.

Random forests. Breiman's bagged ensemble [23] trains independent decision trees on bootstrapped samples and decorrelates them by subsampling features at each split. Two properties matter for agricultural deployment. First, the model degrades gradually when a sensor channel is missing, which is common in field IoT, rather than failing abruptly as a deep network can. Second, mean-decrease-in-impurity and Shapley attributions [24] give an agronomist a clear explanation of why a field was flagged. Random forests remain the default, and often the best-performing method, for yield regression [25], soil organic carbon estimation [26], and multi-temporal crop-type mapping from satellite time series [27].

Support vector machines. The RBF-kernel SVM, $K(\mathbf{x}, \mathbf{x}') = \exp(-\gamma\|\mathbf{x} - \mathbf{x}'\|^2)$, projects inputs into a high-dimensional space and finds the maximum-margin separator. Its main advantage in agriculture is sample efficiency, which matters because labelled hyperspectral pixels are often available only in the hundreds [28, 29]. An SVM can reach a competitive decision boundary with an order of magnitude less data than a comparable deep classifier.

Gradient boosting. XGBoost [30] and LightGBM build an additive ensemble by fitting trees to the negative gradient of a regularised objective, using second-order updates that converge faster than first-order methods. In commercial agriculture, gradient boosting is the method that runs at scale on tabular data: yield forecasting, pest-risk scoring, input-recommendation pipelines, and loss-and-claim modelling. Its prominence in deployment, despite receiving little attention in many recent surveys, shows how far published novelty can be from deployed value.

4.2 Convolutional Neural Networks and Visual Perception

CNNs are responsible for the most visible successes of Agriculture 4.0: in-cab weed detection, fruit counting, leaf-disease apps, and drone-based crop monitoring. They are also the family where the gap between reported accuracy and field accuracy is largest.

Backbones. ResNet [31] solved the depth problem through identity skip connections, $\mathbf{y} = \mathcal{F}(\mathbf{x}, \{W_i\}) + \mathbf{x}$, and remains a standard ImageNet-pre-trained backbone; ImageNet transfer learning is in turn the default recipe for agricultural classification [5]. EfficientNet's compound scaling [32] reduced the per-FLOP accuracy cost of deeper networks; depthwise-separable variants (MobileNetV3, ShuffleNetV2) push inference latency below 10 ms on Jetson-class hardware, which is often what allows a method shown on a desktop GPU to run on a tractor.

Attention modules. Squeeze-and-Excitation [33] recalibrates channels by reducing each feature map to a scalar, passing it through a small MLP, and gating the channel output. CBAM [34] combines channel attention with spatial attention. Both add well under a percent to the parameter count and typically add 1–2 mAP on small-object agricultural detection (rosette weeds, lesion patches, individual fruit), where the discriminative signal occupies only a few pixels among many distractors.

Object detection. The YOLO series [35] performs detection in a single feed-forward pass and covers most agricultural detection tasks well on the speed–accuracy trade-off. YOLOv8 [36] replaces anchor-based heads with anchor-free, decoupled classification and regression branches, which avoids tuning anchor priors to agricultural aspect ratios (elongated wheat heads, near-circular fruits, irregular weed shapes). YOLOv8-N runs at about 80 FPS on a Jetson AGX Xavier, fast enough for the UAV-speed weed detection used in cooperative aerial–ground systems such as Flourish [21].

Semantic segmentation. U-Net [37] preserves fine spatial detail by passing skip connections from the contracting path to the expansive path. It is widely used in agriculture because it works with small datasets: a few hundred pixel-labelled tiles are enough to produce useful canopy and soil masks. SegFormer [38] pairs a hierarchical mix-transformer encoder with an all-MLP decoder, reaching high mIoU at a much lower parameter count, and Mask2Former [39] handles semantic, instance, and panoptic segmentation with a single masked-attention formulation that is robust to the dense, overlapping canopy geometry common in row crops. For small-class targets, loss design matters more than architecture: cross-entropy alone over-fits to background, and a combination of Dice and focal $(1 - p_t)^\gamma \log p_t$ losses is the standard remedy.

Instance segmentation and 3D. Mask R-CNN [40] adds a per-RoI mask branch to Faster R-CNN, which supports per-plant phenotyping (leaf area, stem count) and per-fruit sizing for selective harvesting. Hyperspectral applications use 3D-CNNs that convolve jointly over space and spectrum; hybrid spectral-then-spatial designs reduce cost while retaining spectral discrimination [3].

4.3 Recurrent, Temporal, and State-Space Models

Timing is critical in agriculture. A missed week of pest pressure or a mistimed nitrogen application affects the harvest, not the following week. Sequence models occupy the reasoning stage of the PRA decomposition.

LSTM and bidirectional LSTM. The LSTM cell [41] maintains a cell state c_t and a hidden state h_t updated through gated operations,

$$\begin{aligned} f_t &= \sigma(W_f[h_{t-1}, x_t] + b_f), \\ i_t &= \sigma(W_i[h_{t-1}, x_t] + b_i), \\ c_t &= f_t \odot c_{t-1} + i_t \odot \tanh(W_c[h_{t-1}, x_t] + b_c), \end{aligned}$$

which is enough to retain a season-long record of cumulative growing-degree days while still responding to the past week's weather. LSTMs and Bi-LSTMs are used in most deployed systems for yield prediction from weather data [42], soil-moisture forecasting [43], and phenological staging.

Temporal Convolutional Networks. TCNs [44] stack 1-D dilated causal convolutions whose receptive field grows exponentially with depth, reaching the same effective memory as an LSTM while allowing full parallelisation over time during training. They typically converge several times faster and avoid the gradient problems of backpropagation-through-time. On multivariate weather and soil-moisture benchmarks they match or exceed LSTMs at a fraction of the training cost.

Time-series transformers. Temporal Fusion Transformers [45] combine variable-selection networks, an LSTM-based local stage, and multi-head attention. The attention weights also serve as an interpretable feature-importance map, indicating which historical windows drove a given forecast. PatchTST and similar patch-token methods spread the $O(L^2)$ attention cost across patches, which makes long agricultural sequences such as multi-year hourly IoT data tractable.

State-space models: Mamba. Mamba [46] replaces self-attention with selective state-space models in which the transition matrix depends on the input. Complexity becomes linear in sequence length, which can be the difference between being able to train on a season of hourly data and not being able to. Early agricultural results are encouraging, with LSTM-level accuracy at roughly twice the throughput on long-horizon forecasting, though the literature is still small and direct comparisons are limited.

4.4 Vision Transformers and Hybrid Architectures

The vision transformer [47] treats an image as a sequence of patch tokens and applies stacked self-attention. Its main relevance for agriculture is that the receptive field is global from the first layer. When the discriminative signal is a global pattern, such as a disease pattern that co-varies with field-edge shading or a weed cluster that follows a watercourse, this changes what the model can learn.

Swin and hierarchical variants. Swin Transformer [48] restricts attention to local $M \times M$ windows and shifts the windows between alternate layers to let information flow across them, reducing cost from $O(N^2)$ to $O(NM^2)$. Its hierarchical feature pyramid makes it straightforward to use for downstream detection and segmentation, and its top-1 accuracy advantage over ResNet-50 on plant-disease benchmarks is consistent across recent work.

Data efficiency and DeiT. ViT's need for large training sets is a problem when the agricultural dataset has 5,000 images rather than 5 M. DeiT's distillation-token training [49] transfers convolutional inductive bias from a CNN teacher to the transformer student without full ImageNet-21k pre-training, restoring competitive accuracy on small datasets.

ConvNeXt. ConvNeXt [50] applies the design choices of transformer blocks (large depthwise kernels, inverted bottlenecks, GELU, LayerNorm) to a convolutional backbone. It matches Swin on ImageNet while remaining a standard convolution that compiles well to TensorRT and CoreML, which makes it a common choice for edge deployment in agricultural pipelines where a transformer's irregular memory access is a bottleneck.

4.5 Large Language Models in Agriculture

LLMs do not perceive the field directly, but they have become the standard interface between field data and non-technical farmers, and a small but growing part of the agricultural ML stack now runs through them.

Parameter-efficient adaptation. Full fine-tuning of a 7B–70B parameter LLM on agricultural text is too expensive to be practical. LoRA [51] freezes the pre-trained weight matrix W_0 and learns a low-rank update $\Delta W = BA$ with $r \ll \min(d, k)$, training about 0.1–1 % of the parameters and matching full fine-tuning accuracy in most agricultural adaptation studies. QLoRA extends this by quantising W_0 to 4-bit NormalFloat, allowing 65B-scale fine-tuning on a single 48 GB GPU. AgroGPT and similar domain-adapted models [52] come from this line of work.

Reasoning patterns. Chain-of-thought prompting [53] produces intermediate reasoning steps and improves performance on multi-step agronomic problems (*identify the disease, look up its temperature requirements, compare to the seven-day forecast, recommend the intervention window*). Tree-of-thought and self-consistency variants help when the diagnosis has several branches.

Tool-augmented and agentic patterns. ReAct [54] interleaves reasoning with tool calls, for example to a weather API, a soil database, or a satellite NDVI query, which grounds advisory output in verified, current data. Multi-agent designs divide the work across specialised sub-agents and use an orchestrator to combine their outputs. The main failure mode in agricultural deployment is unreliable tool invocation: an LLM that produces an incorrect API parameter and opens an irrigation valve causes real financial loss, not just a research error.

Vision–language for diagnosis. GPT-4V, LLaVA, and InternVL accept image-and-text prompts, so a farmer can photograph a leaf and ask what is wrong with it in their own language. Their accuracy is limited (see §4.10), but the gain in accessibility is significant. Retrieval-augmented generation [55], which grounds responses in a verified plant-pathology corpus, is the standard way to reduce confident but incorrect pesticide recommendations.

4.6 Graph Neural Networks for Spatial–Temporal Farm Modelling

Many agricultural quantities are better described on a graph than on a grid: neighbouring fields share weather and pests, supply chains form directed graphs from farm to processor to retailer, and rotation history forms a temporal graph across years. Graph Neural Networks [56] learn over $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ by message passing,

$$h_v^{(l+1)} = \phi\left(h_v^{(l)}, \bigoplus_{u \in \mathcal{N}(v)} \psi(h_v^{(l)}, h_u^{(l)}, e_{uv})\right), \tag{1}$$

with ϕ, ψ learnable and $\bigoplus$ a permutation-invariant aggregator. Spatial-graph variants on field-adjacency graphs give better regional pest-pressure estimates than models that treat fields independently, and spatio-temporal GNNs (STGNN, AGCRN) combine graph convolutions with recurrent updates to capture both the direction and the timing of disease outbreaks. Supply-chain GNNs help identify cold-chain bottlenecks and support freshness-aware logistics on farm-to-retail networks.

4.7 Reinforcement Learning for Sequential Farm Decisions

Many agricultural decisions are sequential, with rewards that arrive much later. Irrigating today affects soil moisture for weeks, and a missed fertilisation window cannot be recovered. RL frames this as a Markov Decision Process $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ in which the state captures soil moisture, weather forecast, and crop stage; the action is to irrigate, defer, or apply x kg/ha; and the reward balances yield against input cost.

PPO [57] clips the policy update to keep it within a trust region, which makes it a common choice when small policy errors accumulate over a season. RL irrigation controllers trained inside crop simulators (DSSAT-Gym, AquaCrop-OSPy) report 20–45 % water savings at matched yield, though this figure should be treated with the caution discussed in §5.9. Multi-agent RL extends this to coordinated UAV/UGV fleets, where centralised training with decentralised execution is the common way to transfer a learned coordination policy from simulation to a mixed team of field robots.

4.8 Generative and Diffusion Models for Agricultural Data Synthesis

Synthesis is motivated by the annotation bottleneck in the data layer. Conditional GANs, building on the GAN framework [58], learn paired (condition, image) distributions and can produce, for example, realistic early-stage Fusarium head blight under a specified illumination. Latent diffusion [59, 60] operates in the latent space of a pretrained VAE and is currently the default for text-to-image agricultural augmentation: a prompt such as *"sugar beet seedling with early downy mildew, overhead UAV view, sandy soil, morning light"* produces synthetic training images for rare or geographically narrow conditions. Improvements of 3–8 mAP have been reported on agricultural detection benchmarks when latent-diffusion synthesis replaces classical augmentation, although the trade-off between fidelity and variety is still debated.

4.9 Multi-Modal Foundation Models

CLIP and open-vocabulary detection. CLIP [61] aligns image and text embeddings with a contrastive objective, which supports zero-shot classification by comparing an image embedding against text descriptions of candidate classes. Grounding DINO [62] extends this to open-vocabulary detection, which is useful whenever the weed or pest list cannot be fixed at training time.

DINOv2. DINOv2 [63] self-distils a ViT on a curated corpus and produces dense patch features that transfer well and cluster along semantic boundaries. Domain-specific self-supervised pre-training on unlabelled agricultural imagery improves label efficiency, outperforming ImageNet-supervised initialisation when only a few labels per class are available [64], which matters when each label requires a field visit.

Segment Anything. SAM [65] takes a sparse prompt (point, box, or coarse mask) and returns a pixel-accurate segmentation in tens of milliseconds. As a deployed model it depends on a prompt and is not autonomous, but as an annotation tool it has reduced polygon-labelling cost by close to an order of magnitude, both in reported ablations and in our own use.

ImageBind and unified sensing. ImageBind maps six modalities (image, text, audio, depth, thermal, IMU) to a shared embedding, using image as the anchor. One under-explored use is as the link between sensors in a multi-sensor agricultural pipeline: thermal canopy maps, LiDAR returns, and acoustic soil-moisture data aligned to RGB without needing paired data for every modality combination.

4.10 When Does Each Model Family Deliver?

Listing architectures is the straightforward part of a survey. Determining which ones hold up on a working farm is harder, and it is where the published record and the deployment record diverge. This subsection sets out those differences directly.

Classical ML deserves first-order consideration. A substantial fraction of agricultural "deep learning" results would have been matched or beaten by a well-tuned random forest with engineered features. On structured tabular inputs, RF and gradient boosting still routinely give the best results [25, 26], with advantages a deep network does not have: training in seconds, calibrated probability outputs, built-in handling of missing channels, and attributions (SHAP [24]) that hold up to expert scrutiny. On a new tabular agricultural problem, the sensible first step is to benchmark RF or XGBoost, and the absence of that comparison in many recent papers is a genuine weakness. Classical ML also has clear limits: it cannot use raw imagery or 3D point clouds without manual feature engineering, and its decision boundaries are non-smooth, which generalises poorly under geographic distribution shift.

CNNs: state of the art in-distribution, brittle out of it. YOLO and U-Net derivatives are near the Pareto frontier for in-distribution detection and segmentation. Two failure modes limit their deployment performance. The first is class imbalance. In a working field, about 80 % of weed instances belong to two or three common species, and the rare invasive biotype that most needs early intervention is underrepresented by a factor of 100 to 1000. Without focal loss, class-balanced sampling, or rare-class oversampling, a model recalls common species at above 90 % and rare ones at below 50 %, the opposite of the agronomic priority. The second is geographic shift: a weed detector trained on European sugar-beet imagery typically loses 15–30 mAP when transferred to tropical clay-soil conditions [66, 67]. Few papers report stratified geographic or seasonal splits, which is why reported and deployed numbers differ so consistently.

Temporal models: dependable but data-hungry, and not causal. LSTM and TFT-style architectures consistently outperform statistical baselines on within-distribution forecasting. Three limitations are under-reported. First, data volume: an LSTM yield model typically needs five to seven years of daily history to generalise, and in regions with limited meteorological infrastructure this, rather than model capacity, is the binding constraint. Second, extrapolation: a model trained on 2000–2019 is being asked in 2025 to predict outcomes the historical record does not contain, so yield predictions under unprecedented heat or drought are extrapolations with no calibration guarantee [68, 69]. Third, correlation rather than causation: a model that associates early-season NDVI with final yield may be relying on variety choice or regional management, and under a change in those confounders its predictions fail in ways that benchmark error does not reveal.

Vision transformers: powerful, often disqualified by cost. ViT-B ($\sim$86 M parameters, $\sim$55 GFLOPs) and YOLOv8-N ($\sim$7 M parameters, $\sim$1.8 GFLOPs) belong to different deployment regimes. For smallholder smartphone apps, solar-powered field nodes, and UAV onboard inference, the difference is decisive unless the ViT is heavily quantised or distilled. The few-shot benefit of pretrained ViTs is real but depends on the pretraining domain being close to the target, and ImageNet-21k priors transfer poorly to hyperspectral or thermal agricultural data.

LLMs: unprecedented accessibility, unsafe numerics. A voice-driven LLM advisor is a major improvement in accessibility for a farmer who cannot use a precision-agriculture dashboard. The deployment risk is that LLMs produce confident but incorrect quantitative agronomy: pesticide rates, tank-mix compatibility, and variety-specific emergence thresholds. RAG [55] grounded on verified extension material reduces this substantially, but corpus quality varies sharply by region. For North American and EU crops the grounding corpus is adequate; for many South Asian and sub-Saharan African staples it is not, and an ungrounded LLM in those settings is a genuine safety concern.

Foundation models: zero-shot is a triage tool, not a diagnostic one. CLIP zero-shot disease classification reaches 50–65 % on standard benchmarks, whereas fine-tuned supervised models reach over 90 %. The useful deployment role of foundation models is therefore not zero-shot inference but sample-efficient fine-tuning (DINOv2) and faster annotation (SAM, Grounding DINO), where they mainly reduce development cost rather than improve headline accuracy.

5 Application Domains

We cover eight application domains that together account for most of the field-level decisions on a working farm. Figure 5 shows the mapping between model families and domains, Table 3 lists the most widely used datasets, and Table 4 lists representative methods, the dataset each was evaluated on, and the reported metric.

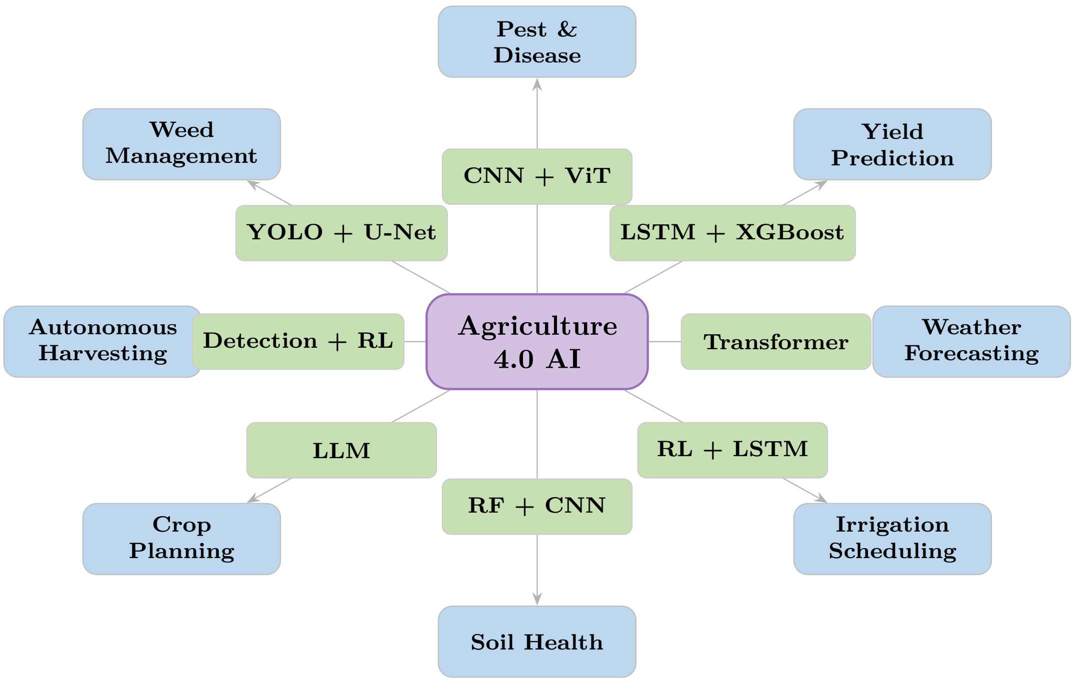

Figure 5. Application domains mapped to their dominant model families.

Table 3. Benchmark datasets used across the Agriculture 4.0 application domains.

Dataset	Primary Task	Size	Modality	Public	Representative Models
PlantVillage [5]	Disease classification	54,306 img	RGB leaf photos	Yes	ResNet, VGG, MobileNet
WeedNet [66]	Weed segmentation	562 images	RGB + NIR (UAV)	Yes	SegNet, FCN
CWFID	Weed detection	60 annotated img	RGB (field)	Yes	CNN, SSD
IP102 [70]	Pest classification	75,222 images	RGB	Yes	CNN, ViT, ResNet
Global Wheat Head [71]	Spike detection	4,700 imgs; 193,634 heads	RGB (ground + UAV)	Yes	YOLO, Faster R-CNN
Sentinel-2 (ESA)	Crop mapping; yield	Global time-series	12-band MSI	Yes	LSTM, CNN, RF
LUCAS Soil (EU JRC)	SOC & property est.	22,694 samples	NIR spectra + lab	Yes	PLS, Random Forest, 1D-CNN
DeepSat (SAT-4/SAT-6)	Land-cover classification	500,000 patches	4-band aerial	Yes	CNN, SVM
iNaturalist (plant subset)	Species classification	>200K images	RGB	Yes	EfficientNet, ViT
OpenFarm / FarmHack	Crop planning / advisory	Text & structured	Agronomic records	Yes	LLM fine-tuning, RAG

Table 4. Representative methods per application domain.

Domain	Reference	Model / Approach	Dataset	Metric	Key Result / Note
Weed Detection	Sa et al. [66]	SegNet (RGB+NIR)	WeedNet	F1	~0.80 (0.78 AUC); multispectral gain
	Milioto et al. [67]	FCN (real-time)	Sugar-beet field	FPS	>10 FPS at vehicle speed
	Pretto et al. [21]	UAV–UGV spraying	Sugar-beet field	Qual.	Aerial map drives selective spray
Pest & Disease	Mohanty et al. [5]	AlexNet / GoogLeNet	PlantVillage	Acc.	99.35 % (lab conditions)
	Ferentinos [6]	VGG-16	87K-image set	Acc.	99.53 % top model
	Wu et al. [70]	FRCNN/SSD baselines	IP102	—	75K imgs, 102 classes (long-tail)
Yield Prediction	Nevavuori et al. [68]	Deep CNN	RGB/NDVI UAV seq.	MAE	484 kg/ha (~8.8% MAPE); RGB>NDVI
	Pantazi et al. [72]	Supervised SOM/ANN	Soil vis–NIR + satellite	R^2	High-accuracy wheat yield map
	You et al. [69]	CNN/LSTM + GP	MODIS	RMSE	County-level soybean (US)
Weather Forecast	Salman et al. [73]	RNN/CRBM/ConvNet	Station records	MAE	DL station forecast
	Bi et al. [74]	Pangu-Weather (Transf.)	ERA5 reanalysis	RMSE	NWP parity at 7-day lead
Irrigation Scheduling	Liu et al. [75]	Deep RL (review: DQN/PPO)	Crop simulators	Water	20–45% savings reported (sim.)
	Mohamed et al. [76]	IoT sensor-based irrigation control	IoT field trials	Water	47 % less water, 43 % higher yield
Soil Health	Viscarra Rossel et al. [26]	vis–NIR spectroscopy + data mining	Global soil library	R^2	SOC, clay & CEC prediction
	Hengl et al. [77]	Random Forest + GBM	MODIS/terrain covariates	—	Global SOC & soil-class mapping
Crop Planning	Awais et al. [52]	AgroGPT vision-language model	Ag. databases	Eval	Multimodal agriculture Q&A

mAP = mean average precision; mIoU = mean intersection over union; RMSE = root mean squared error; MAE = mean absolute error.

5.1 Weed Management

Weeds are the largest single source of pre-harvest yield loss across most crops, with estimated global potential losses of around 34 % [78]. Broadcast herbicide application has controlled them for half a century, but at the cost of ecosystem load and the rapid emergence of resistant biotypes.

The now-dominant computer-vision approach is real-time CNN detection driving patch-level rather than broadcast spraying. Sa et al. [66] showed the value of NIR augmentation in a SegNet-style architecture, and Milioto et al. [67] brought FCN-based crop/weed discrimination to real time at vehicle speed on sugar beet. The UAV/UGV cooperative approach used in the Flourish system [21] pairs an aerial multispectral survey with a ground rover for selective spraying, reducing chemical input relative to a broadcast pass without reducing coverage. Several problems remain unsolved, including resistance identification, generalisation to novel species, and late-stage canopy occlusion.

5.2 Pest and Disease Management

The PlantVillage work [5, 22] showed that CNN classification can reach above 99 % accuracy on isolated leaves under controlled conditions. Ferentinos [6] extended this to 87,000 images across 58 classes with similar reported numbers. Later in-field work showed that these numbers do not transfer directly: complex backgrounds, variable illumination, and partial occlusion cause a substantial accuracy drop on unconstrained photographs [79]. For arthropod pests, the IP102 benchmark [70] is a more realistic 102-class target, and the strongest CNN baselines in the original study reach only about 49 % top-1 accuracy, which leaves large room for improvement on rare classes. The practical lesson is that classification results from leaf-isolation datasets should be treated as upper bounds on field performance rather than as field performance.

5.3 Yield Prediction

Yield prediction combines meteorology, soil profiles, management records, satellite indices, and historical yield maps over time scales of weeks to months. Deep-learning methods are now common: CNNs support UAV-image yield estimation [68], and CNN/LSTM models coupled with Gaussian processes are used on county-level remote-sensing benchmarks [69]. Attention-augmented LSTMs and Temporal Fusion Transformers add interpretability by exposing weights over historical windows. Wheat-specific work using supervised SOM/ANN hybrids on online soil-spectroscopy and satellite inputs produces accurate regional wheat-yield maps [72]. Within-field yield variability, which is what matters for variable-rate decisions, is considerably harder to predict than the regional averages most papers report.

5.4 Weather and Climate Forecasting

There are two distinct settings. Short-range, station-level forecasting from local sensor data benefits from deep architectures such as recurrent, convolutional, and conditional restricted Boltzmann machine networks [73]. At the global scale, transformer-based models now approach the accuracy of numerical weather prediction: Pangu-Weather and GraphCast [74, 80] reach operational accuracy at 7–14 day lead times against ERA5 reanalysis. The open question for agriculture is downscaling. Regional and field-scale forecasts derived from these global models are still at an early stage, and producing them cheaply is a prerequisite for connecting weather prediction to irrigation scheduling.

5.5 Irrigation Scheduling

Roughly 70 % of global freshwater use goes to agriculture, so even small percentage improvements have large absolute effects. Supervised models trained on in-field sensor data can already reproduce much of an expert agronomist's irrigation decision-making [81]. A recent review by Liu et al. [75] of deep-RL irrigation controllers reports water savings broadly in the 20–45 % range against calendar baselines at matched yield in simulation, with results that depend heavily on rainfall variability and soil assumptions. Mohamed et al. [76] describe an IoT-driven smart irrigation system with real-time sensor monitoring and control, reporting a 47 % reduction in water use and a 43 % increase in yield. A common architecture pairs a soil-moisture predictor (LSTM, TCN, or Mamba) with a controller (model-predictive control or RL) and runs it inside a digital twin (§6.3). The main constraint on deployment is the sim-to-real gap (§5.9).

5.6 Soil Health Monitoring

Soil organic carbon, nitrogen, pH, bulk density, and texture underpin fertility management and are increasingly relevant to carbon-credit accounting. Wet-chemistry analysis is accurate but expensive and slow.

Spectroscopy-based estimation. Near-infrared (NIR) and mid-infrared (MIR) spectroscopy on dried, ground samples is the standard operational method [26], with PLS regression as the usual baseline. On heterogeneous continental-scale datasets (LUCAS is the standard example), 1D-CNNs applied directly to spectral profiles outperform PLS by learning non-linear absorption-feature mappings without manual band selection, and their filters often recover physically meaningful absorption peaks such as the ∼2,100 nm C–H band.

Remote sensing and spatial mapping. UAV hyperspectral imagers extend spectroscopy to whole-field surface mapping, and satellite missions such as DESIS and PRISMA extend it to regional scale. Bare-soil composites from Sentinel-2, combined with terrain attributes (slope, curvature, topographic wetness index) from DEMs, feed RF and CNN models for soil-type and SOC mapping at regional-to-global scale [77]. EMI sensors (apparent EC) add sub-surface information, and combining them with spectral data improves clay-content and salinity prediction.

Emerging deep models. Attention-based architectures over multi-source soil data (spectra, terrain, satellite time series) improve accuracy further. Treating wavelength positions as token indices in a transformer is a natural way to capture long-range spectral dependencies. GNNs over spatially structured monitoring networks aggregate field-level observations while accounting for spatial autocorrelation.

5.7 Crop Planning and Decision Support

Rotation, variety choice, input allocation, and planting dates form a multi-factor decision space that is too large for manual rule sets.

LLM-based advisory. AgroGPT and similar systems [52] fine-tune general-purpose LLMs on agricultural extension corpora to produce natural-language plans for planting, disease management, and soil amendments. RAG [55] grounded on regional databases, together with chain-of-thought prompting [53], addresses most of the quantitative-hallucination problem in well-documented regions. The risk remains in regions where the grounding corpus is thin.

Optimisation hybrids. A promising emerging pattern pairs an LLM (for eliciting constraints in natural language and specifying the problem) with a classical solver (for global optimality on the resulting linear, mixed-integer, or multi-objective program). This separates the component that must be easy to interact with from the component that must be correct.

Multi-crop polyculture. Smallholder polyculture systems, with intercropping and complex interactions between species, are best modelled as partially observed Markov decision processes. RL agents trained inside crop simulators (DSSAT, APSIM) can learn planting policies that balance short-term income against long-term soil and biodiversity outcomes.

5.8 Autonomous Harvesting

Harvesting is the most mechanically demanding application, requiring precise manipulation under high variability in fruit position, occlusion, and surface compliance. A common architecture pairs YOLO or Mask R-CNN with stereo or depth perception to localise the fruit [10, 82], together with model-based or learned inverse kinematics for the arm trajectory. Xiong et al. [83] demonstrated an end-to-end pick-and-place system for strawberries. Commercial platforms (Dogtooth, FFRobotics, asparagus harvesters) are close to viability, with cost and reliability, rather than perception accuracy, as the binding constraints.

5.9 The Lab-to-Field Gap

The methods above are drawn from the peer-reviewed record. Understanding how the published numbers behave under deployment is essential for judging the actual maturity of the field, and this subsection sets out those differences directly.

The benchmark validity problem. The most-cited result in agricultural disease detection, 99.35 % accuracy on PlantVillage [5], was obtained on isolated leaves photographed against uniform backgrounds under controlled lighting. When the same models are re-evaluated on smartphone images of real diseased plants under variable illumination and background clutter, accuracy typically falls to 60–75 %, a drop of 25–40 percentage points. This is not an exception; it is the usual outcome across nearly every agricultural CV benchmark that has been re-tested in field conditions. The cause is a mismatch in evaluation protocol. Most papers split train and test by random sampling within the same acquisition session, which leaks spatial and temporal correlations. Site-held-out, year-held-out, or geographically stratified splits give numbers that actually predict deployment performance, and few papers use them.

Where precision weed management breaks down. Four failure modes recur in farm-scale deployments. *Canopy closure*: when weed density is high, plants overlap and detectors trained on isolated instances cannot separate them. *Growth-stage mismatch*: a model trained on seedling-stage rosettes does not work on bolted plants, which is a problem because bolted plants cause the most damage. *Resistance identification*: herbicide resistance is biochemical rather than morphological, so no vision model can identify a resistant biotype from imagery alone, even though the management decision depends on it. *Novel species*: a model with no training examples of an invasive species cannot detect it, and the retraining loop needed to handle a new invasion is not supported by most production platforms.

Yield prediction: the calibration gap. Reported $R^2 > 0.85$ values are typical, but they are usually computed against test years that are climatically similar to the training period. The economically important task, predicting a drought year unlike any in the training data, is exactly the case where the model extrapolates beyond its support, and the calibration error there is far larger than the reported number suggests. Separately, the gap between county-level and within-field accuracy is large, because spatial averaging at county scale removes variation that a variable-rate decision needs.

Reinforcement learning for irrigation: simulation-to-real transfer. The simulator-only studies that report around 30 % water savings rely on homogeneity assumptions (uniform soil, no preferential flow, parametric ET) that no real soil satisfies. Field deployments regularly report controllers that over-irrigate clay, because drainage lags the simulator, or under-irrigate sand, because preferential flow is not modelled. Multi-season field trials with control plots remain rare, so the published savings should be read as upper bounds.

Deployment cost versus published model complexity. There is a clear selection bias in the literature: complex, high-accuracy models are published more often than the simpler models actually deployed at scale. The platforms that operate across millions of hectares (John Deere Operations Center, Climate FieldView, Granular) run gradient-boosted trees, shallow CNNs, and LSTMs rather than ViT-L or 70B LLMs. The reason is economic: at per-field, per-day inference rates over that area, the compute cost of large models is prohibitive, and their marginal accuracy gain on structured sensor data is too small to justify. A realistic survey has to distinguish between what wins a benchmark and what is economically viable at agricultural scale.

Long-tail performance. Agricultural class distributions are strongly long-tailed: common species and conditions dominate annotation counts, while rare but important ones are scarce. Top-1 accuracy and aggregate mAP hide this. A classifier with 95 % overall accuracy may be at chance on the rare disease classes that most need early intervention. Per-class reporting and explicit rare-class evaluation should be standard in agricultural ML publications, and at present they are not.

6 Future Directions

This section describes the research directions we consider most important for the next phase of agricultural AI represented in Figure 6. We keep them separate from the deployment barriers in §7: the distinction is whether the open problem is mainly about building new capability (this section) or about making existing capability work in practice (the next).

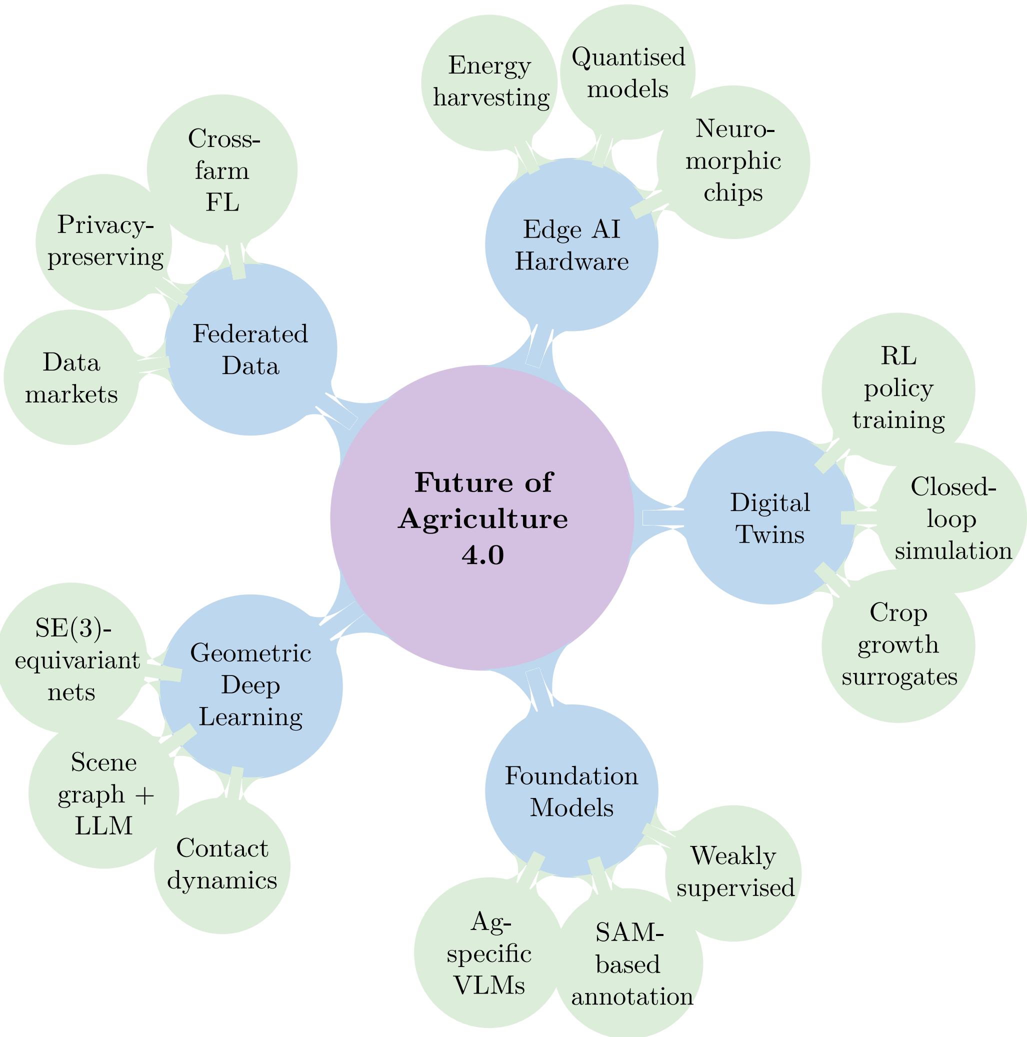

Figure 6. Research directions for the next phase of agricultural AI.

6.1 Geometric Deep Learning and Physical AI

CNNs and ViTs operate on grid-structured pixels and have no built-in representation of the 3D geometry that governs robot–plant contact. Geometric deep learning [84] extends learning to non-Euclidean domains such as point clouds, meshes, and graphs. SE(3)-equivariant networks [85] build invariance to 3D rotation and translation into the architecture, so that rotating a fruit-tree LiDAR scan produces a correspondingly rotated grasp prediction without pose normalisation or augmentation. The benefit for agriculture is direct learning of grasp points and approach trajectories on raw point clouds, without voxelisation or projection artefacts.

The broader *Physical AI* approach pairs geometric deep learning (for geometry and contact) with LLMs (for task specification). A high-level command (*"harvest the ripe apples in this row"*) is decomposed into a sequence of geometric sub-goals (detect, approach, grasp, retract), each executed by the physical arm. Contact-rich interactions, such as compliant grasping of soft produce, navigation through dense canopy, and micro-injection of nematicides, are the main open problem and the case where CNN-only perception is clearly insufficient.

The Flourish UAV/UGV system [21] is an early example of this approach: aerial CNN inference produces geo-referenced task specifications that a ground robot executes. Replacing the hand-designed coordination layer with a learned, geometry-aware MARL policy is the natural next step, and it requires simulation environments (§6.3) that model contact and deformation realistically.

6.2 Agricultural Foundation Models

General-purpose vision foundation models are consistently suboptimal for agriculture because their pretraining data contains almost no canopy texture, soil spectral signatures, oblique UAV viewpoints, or early-stage-disease leaf morphology. An agriculture-specific foundation model pretrained on a large, geographically diverse field corpus would reduce downstream fine-tuning requirements from thousands of labels to tens.

Three complementary strategies are plausible. First, self-supervised pretraining (masked autoencoders, DINOv2-style self-distillation) on unlabelled multispectral satellite archives (Sentinel-2, MODIS, Planet) and UAV imagery across regions, crops, and growth stages. Second, semi-automatic annotation pipelines that combine SAM [65] mask generation, Grounding DINO [62] open-vocabulary localisation, and active-learning sample selection to reduce labelling cost by an order of magnitude. Third, synthetic generation from physics-based crop-growth simulators (AgroSim, OpenPlant3D, Helios) for rare disease phenotypes and extreme-weather conditions that real data does not cover.

A unified agricultural foundation model would ideally accept all sensing modalities (RGB, multispectral, hyperspectral, thermal, LiDAR, SAR) through a shared encoder, with lightweight adapter heads for classification, detection, segmentation, and time-series tasks. Building it is mainly an engineering and data-collection effort rather than a methods problem, and the main obstacle is coordination, namely which institutions contribute, under what licensing, and with what compute, rather than research.

6.3 Digital Twins and Simulation-Assisted Autonomy

A field-level digital twin [86] is a continuously synchronised virtual replica of a farming system that integrates satellite-derived canopy state (NDVI, LAI, chlorophyll), soil sensor streams, high-resolution weather forecasts, and the current state of mechanistic crop models (DSSAT, APSIM, WOFOST, AquaCrop). Updated sub-hourly from IoT telemetry, it provides a low-cost way to run what-if analyses: *"if I withhold irrigation for 48 hours during the forecast dry window, what happens to soil-moisture deficit and end-of-season yield?"*

Neural surrogates such as physics-informed neural networks and deep Gaussian processes can replace the expensive mechanistic simulator inside the optimisation loop, reducing multi-day forward simulations to millisecond forward passes and enabling receding-horizon model-predictive control of irrigation, fertilisation, and harvest scheduling.

The most important use of a digital twin, however, is as a safe environment for RL training [57]. A PPO policy that maximises season-long yield inside the twin can then be deployed on the physical farm without exposing it to the cost of in-field exploration. Domain randomisation (varying soil hydraulic parameters, weather variability, and sensor noise during training), together with adversarial domain alignment against real telemetry, is the standard sim-to-real bridge.

Aggregating per-farm twins into a regional twin enables regional yield forecasting at scale and early identification of spatially correlated pest-outbreak risk before it appears in the field.

6.4 Agentic AI Systems for Autonomous Farm Management

LLM-based agentic systems [54] go beyond static advisory output toward goal-directed autonomous operation: they perceive farm state from IoT, satellite, and weather APIs; reason in natural language; select actions from a tool library (call the irrigation controller, schedule a drone survey, order inputs, file a claim); and then execute and iterate. The tool library can be extended without retraining, so adding access to a carbon-credit marketplace is a tool specification rather than a model update.

Multi-agent architectures divide the work across specialised sub-agents (soil health, disease surveillance, market intelligence), with an orchestrator that combines their outputs into a single plan. Coordination patterns include shared blackboards, hierarchical planning, and debate-based verification, in which a critic agent checks the plan for logical consistency and regulatory compliance before execution.

The main engineering risk is reliable tool invocation under load. The LLM must format API calls correctly, handle unexpected return values, and maintain consistent goal state across multi-day execution traces. Formal verification of agent traces and structured-output schemas (Pydantic, JSON Schema) are the current best practice for safety-relevant agricultural agents.

6.5 Continual and Lifelong Learning

Distribution shift is inherent in agriculture: weed populations, climate, pest pressure, and even soil chemistry change from year to year. A weed model trained in Nebraska in 2022 degrades quietly when applied to Texas weeds in 2025. Continual learning [87] aims to address this by acquiring new knowledge without catastrophically forgetting the old.

Elastic Weight Consolidation penalises changes to parameters important for previous tasks, with importance estimated from the Fisher information matrix. Progressive Neural Networks add lateral columns for each new task, preserving the earlier ones. Dark Experience Replay keeps a small episodic memory of previous samples to interleave with new training. For year-round agricultural robots, online continual learning, which updates from the inference stream without separate retraining sessions, is what allows adaptation to within-season change without redeployment.

6.6 Edge AI and Neuromorphic Hardware

The next generation of edge inference targets 1–10 W power budgets with TOPS-level throughput. Hailo-8M delivers 26 TOPS at 2.5 W, Qualcomm RB5 provides 15 TOPS at 5 W, and Intel Loihi [88] demonstrates on-chip learning for spiking workloads at sub-watt power.

Software-side compression has kept pace. Post-training INT8/INT4 quantisation and structured filter pruning routinely halve the model footprint at under 2 % accuracy cost. Knowledge distillation [89] trains a compact student to match a teacher's softened output distribution, $\mathcal{L}_{\mathrm{KD}} = T^2\,\mathrm{KL}(\sigma(z_s/T)\|\sigma(z_t/T))$, with T the temperature. A YOLOv8-N distilled from YOLOv8-L typically retains about 95 % of the teacher's mAP at about 40 % of the latency, which can be the difference between feasible and infeasible on a sub-15 W platform.

Event cameras (iniVation DAVIS346, Prophesee EVK4) emit per-pixel brightness-change events rather than frames, giving microsecond temporal resolution at kilopixel spatial resolution. Because most agricultural scenes are static, they transmit two to three orders of magnitude fewer bytes per second than frame cameras. Spiking networks on neuromorphic processors such as Loihi can process these events for high-speed pest tracking and pick-point detection at sub-watt power, which makes always-on perimeter monitoring feasible without regular battery changes.

6.7 Federated and Privacy-Preserving Farm Data Ecosystems

Per-farm datasets are too narrow to support a robust global model, while pooling data forces farms to share commercially sensitive management records. Federated learning [4] offers a third option. In FedAvg, a server broadcasts the current global model θ^t to a random set of K farms; each farm runs E local epochs; and the server aggregates $\theta^{t+1} = \sum_k (n_k/n)\theta_k^{t+1}$, weighted by local dataset size.

Differential privacy limits information leakage by clipping per-sample gradients to norm C and adding calibrated Gaussian noise, $\tilde{g} = g/\max(1, \|g\|/C) + \mathcal{N}(0, \sigma^2 C^2 I)$, giving (ε, δ)-DP guarantees that bound how much any single farm's records can leak. Personalised variants (FedProx, SCAFFOLD, pFedMe) keep a locally adapted model for each farm while sharing a global backbone.

The underdeveloped part is governance rather than algorithms. Federated data markets, in which smart contracts compensate farms in proportion to their data's contribution (Shapley-value federated attribution), are a natural next step toward the data volume needed for agricultural foundation models. Cross-silo FL across research institutions, agribusinesses, and government agencies is the most likely source of the geo-diverse multi-modal datasets required.

7 Challenges and Open Problems

Section 6 described capabilities to build. This section describes structural barriers that prevent existing capabilities from being used. The two are easy to conflate but worth separating: building a better detector and getting an existing detector to work safely on a Kenyan maize plot are different problems.

Domain shift and geographic generalisation. A weed detector trained on sandy-soil Iowa soybean and deployed to clay-loam Kenyan maize encounters different soil colour, weed species, growth stages, illumination, and viewpoints at the same time. Adversarial feature alignment, test-time adaptation of batch-normalisation statistics, and domain-randomised pretraining each address part of the problem and each risk negative transfer. The more fundamental fix is evaluation discipline: multi-environment test protocols and geographically stratified, open datasets that span global cropping diversity. How quickly the lab-to-field gap closes will depend on the community's willingness to publish the weaker numbers that harder splits produce.

Calibrated uncertainty and safety. For actions such as applying a herbicide that might damage a neighbouring crop, or commanding an arm to grasp a fruit, the model's confidence is as important to the decision as its prediction. Standard softmax outputs are systematically overconfident. Bayesian deep learning (MC dropout, deep ensembles, Bayesian NNs) produces calibrated intervals, and conformal prediction provides distribution-free coverage guarantees ($\Pr[y \in \hat{C}(x)] \geq 1 - \alpha$) that can be applied as a black-box wrapper. Both are underused in deployed agricultural AI, with predictable consequences when deployment moves outside the training distribution.

Interpretability and farmer trust. An incorrect neural decision that drives a chemical or mechanical action has consequences that require a higher standard of interpretability than typical computer-vision deployments. Grad-CAM and its variants produce coarse saliency maps but do not explain why the highlighted region is discriminative. SHAP [24] provides principled feature-level attributions and is a good default for tabular agronomic predictions. For LLM advisory output, chain-of-thought explanations [53] appear to support farmer trust and uptake in early field studies, though the human-factors literature is still limited. The remaining gap is between technical explainability and an explanation that a non-technical farmer can act on.

Cost and accessibility in smallholder contexts. A commercial UAV costs $5,000–50,000, a UGV $20,000–200,000, and a dense IoT sensor network $5,000–20,000 per 10 ha. This equipment is affordable for large commercial operations in high-income economies but out of reach for the roughly 500 million smallholder farms that produce more than 70 % of food in developing countries. Three approaches help: smartphone-deployable models (MobileNetV3, EfficientNet-Lite) for camera-based diagnosis; shared-infrastructure services (drone-as-a-service, cooperative sensor pools) that spread hardware cost across a group; and satellite-based analytics (Sentinel-2, Planet) that require no farmer-side hardware. AI for smallholder settings also has to tolerate intermittent connectivity, support multilingual voice interfaces, and produce useful predictions from sparse, informal records.

Regulatory environment for autonomous robots. Beyond-visual-line-of-sight UAV operation and fully autonomous ground robots on land near public areas are still lightly regulated in most jurisdictions. EU 2019/947 defines operational categories by risk, but agricultural-specific BVLOS provisions are still evolving. Liability rules for AI-induced crop damage, for example a robot that misclassifies a crop row as a weed zone and applies herbicide, are poorly defined in both common-law and civil-law systems. Certification standards (ISO/IEC TR 5469, SOTIF) are being adapted from automotive and aviation but have not yet produced an agricultural standard. This uncertainty reduces capital investment in commercial Agriculture 4.0 platforms regardless of technical readiness.

Data sovereignty and competitive sensitivity. Geo-referenced yield maps, input records, soil profiles, and financial records are commercially sensitive, and in concentrated commodity markets they reveal competitive position. Farmers generate the data, but the models trained on it are usually proprietary to the platform. Federated learning and on-device inference address part of this tension technically. At the governance level, farm-data codes of practice (the U.S. Ag Data Transparency Evaluator framework, EU farm-data sharing initiatives) and open data licensing are the necessary complement to those technical mechanisms.

Robustness under adverse deployment conditions. Agricultural robots regularly operate under conditions largely absent from their training data: midday sun saturating sensors, overcast light flattening spectral contrast, deep canopy shadow, rain, dust, wind-induced motion blur, mud on lenses, and GPS multipath in orchard alleys causing metre-scale localisation error. Augmentation helps, but systematic adverse-condition test sets, the agricultural equivalent of automotive corner-case benchmarks, are largely missing. Producing and standardising them is one of the more concrete near-term contributions the research community could make.

8 Conclusion

What separates Agriculture 4.0 from precision agriculture is not better sensors or better models but the closure of the loop: sense, decide, act, and sense again, without a human necessarily in the chain. This survey is mainly about the decision layer that closes that loop.

We reviewed the four agricultural revolutions, divided the modern stack into hardware, connectivity, data, and intelligence layers, and compared eleven model families on the seven dimensions that determine which one runs in production. We mapped those families onto eight application domains and tabulated their representative methods, datasets, and reported metrics. Two analytical sections, one on what each model family delivers in practice and one on the lab-to-field gap that benchmark numbers hide, connect the survey to deployment. We closed with seven research frontiers and seven deployment barriers, kept separate.

The direction we are most confident about is the combination of geometry-aware perception with language-driven goal specification: a *Physical AI* architecture in which an LLM decomposes a high-level agronomic objective into geometric sub-goals that an SE(3)-equivariant controller executes. This pairing fits agricultural robotics well, providing physically grounded perception, flexible task specification, and robust operation in unstructured outdoor environments. Aerial intelligence coordinating ground action, as in the Flourish system [21], is the architecture we expect to define the next decade of autonomous farms.

The main remaining obstacles, however, are not primarily algorithmic. They are data (geo-diverse, federated, well-licensed), hardware (cheaper edge inference, better power budgets), and governance (regulatory clarity, data ownership, interpretability standards). Progress on these will determine how quickly the capabilities surveyed here reach the parts of global food production that need them most.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Food and Agriculture Organization of the United Nations. The future of food and agriculture: Trends and challenges. Technical report, FAO, Rome, 2017. Available online: https://openknowledge.fao.org/server/api/core/bitstreams/2e90c833-8e84-46f2-a675-ea2d7afa4e24/content (accessed on May 2026).

[2] Klaus Schwab. *The Fourth Industrial Revolution*. World Economic Forum, Geneva, 2016. Available online: https://www.weforum.org/stories/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/ (accessed on May 2026).

[3] Telmo Adão, Jonáš Hruška, Luís Pádua, José Bessa, Emanuel Peres, Raul Morais, and Joaquim João Sousa. Hyperspectral imaging: A review on UAV-based sensors, data processing and applications for agriculture and forestry. *Remote Sensing*, 9(11):1110, 2017. doi: 10.3390/rs9111110.

[4] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR, 2017. Available online: https://proceedings.mlr.press/v54/mcmahan17a.html (accessed on May 2026).

[5] Sharada P. Mohanty, David P. Hughes, and Marcel Salathé. Using deep learning for image-based plant disease detection. *Frontiers in Plant Science*, 7:1419, 2016. doi: 10.3389/fpls.2016.01419.

[6] Konstantinos P. Ferentinos. Deep learning models for plant disease detection and diagnosis. *Computers and Electronics in Agriculture*, 145:311–318, 2018. doi: 10.1016/j.compag.2018.01.009.

[7] Konstantinos G. Liakos, Patrizia Busato, Dimitrios Moshou, Simon Pearson, and Dionysis Bochtis. Machine learning in agriculture: A review. *Sensors*, 18(8):2674, 2018. doi: 10.3390/s18082674.

[8] Andreas Kamilaris and Francesc X. Prenafeta-Boldú. Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, 147:70–90, 2018. doi: 10.1016/j.compag.2018.02.016.

[9] Diego Inácio Patrício and Rafael Rieder. Computer vision and artificial intelligence in precision agriculture for grain crops: A systematic review. *Computers and Electronics in Agriculture*, 153:69–81, 2018. doi: 10.1016/j.compag.2018.08.001.

[10] Anand Koirala, Kerry B. Walsh, Zhenglin Wang, and Cheryl McCarthy. Deep learning – method overview and review of use for fruit detection and yield estimation. *Computers and Electronics in Agriculture*, 162:219–234, 2019. doi: 10.1016/j.compag.2019.04.017.

[11] Muhammad Hammad Saleem, Johan Potgieter, and Khalid Mahmood Arif. Plant disease detection and classification by deep learning. *Plants*, 8(11):468, 2019. doi: 10.3390/plants8110468.

[12] Spyros Fountas, Nikos Mylonas, Ioannis Malounas, Efthymios Rodias, Christoph Hellmann Santos, and Erik Pekkeriet. Agricultural robotics for field operations. *Sensors*, 20(9):2672, 2020. doi: 10.3390/s20092672.

[13] Lefteris Benos, Aristotelis C. Tagarakis, Georgios Dolias, Remigio Berruto, Dimitrios Kateris, and Dionysis Bochtis. Machine learning in agriculture: A comprehensive updated review. *Sensors*, 21(11):3758, 2021. doi: 10.3390/s21113758.

[14] Alan L. Olmstead and Paul W. Rhode. Reshaping the landscape: The impact and diffusion of the tractor in american agriculture, 1910–1960. *The Journal of Economic History*, 61(3):663–698, 2001. doi: 10.1017/S0022050701030042.

[15] Robert E. Evenson and Douglas Gollin. Assessing the impact of the green revolution, 1960 to 2000. *Science*, 300(5620):758–762, 2003. doi: 10.1126/science.1078710.

[16] John V. Stafford. Implementing precision agriculture in the 21st century. *Journal of Agricultural Engineering Research*, 76(3):267–275, 2000. doi: 10.1006/jaer.2000.0577.

[17] Chunhua Zhang and John M. Kovacs. The application of small unmanned aerial systems for precision agriculture: A review. *Precision Agriculture*, 13(6):693–712, 2012. doi: 10.1007/s11119-012-9274-5.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, pages 1097–1105, 2012. Available online: https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks (accessed on May 2026).

[19] Antonis Tzounis, Nikolaos Katsoulas, Thomas Bartzanas, and Constantinos Kittas. Internet of things in agriculture, recent advances and future challenges. *Biosystems Engineering*, 164:31–48, 2017. doi: 10.1016/j.biosystemseng.2017.09.007.

[20] Lele Xu, Zhigang Cao, Meng Tian, Lin Feng, Yan Wang, Fan Zhang, and Boyang Liu. A review of intelligence for agricultural robots and field machinery. *Smart Agricultural Technology*, 1:100011, 2021. doi: 10.1016/j.atech.2021.100011.

[21] Alberto Pretto, Stéphanie Aravecchia, Wolfram Burgard, Nived Chebrolu, Christian Dornhege, Tillmann Falck, Freya Fleckenstein, Alessandra Fontenla, Marco Imperoli, Raghav Khanna, Frank Liebisch, Philipp Lottes, Andres Milioto, Daniele Nardi, Sandro Nardi, Johannes Pfeifer, Marija Popović, Ciro Potena, Cédric Pradalier, Elisa Rothacker-Feder, Inkyu Sa, Alexander Schaefer, Roland Siegwart, Cyrill Stachniss, Achim Walter, Wera Winterhalter, Xiaolong Wu, and Juan Nieto. Building an aerial–ground robotics system for precision farming: An adaptable solution. *IEEE Robotics & Automation Magazine*, 28(3):29–49, 2021. doi: 10.1109/MRA.2020.3012492.

[22] David P. Hughes and Marcel Salathé. An open access repository of images on plant health to enable the development of mobile disease diagnostics. *arXiv preprint arXiv:1511.08060*, 2015. doi: 10.48550/arXiv.1511.08060.

[23] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

[24] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017. Available online: https://arxiv.org/abs/1705.07874 (accessed on May 2026).

[25] Jig Han Jeong, Jonathan P. Resop, Nathaniel D. Mueller, David H. Fleisher, Kyungdahm Yun, Ethan E. Butler, Dennis J. Timlin, Kyo-Moon Shim, James S. Gerber, Vangimalla R. Reddy, and Soo-Hyung Kim. Random forests for global and regional crop yield predictions. *PLOS ONE*, 11(6):e0156571, 2016. doi: 10.1371/journal.pone.0156571.

[26] Raphael A. Viscarra Rossel, Thorsten Behrens, Eyal Ben-Dor, David J. Brown, José A. M. Demattê, Keith D. Shepherd, Zhou Shi, Bo Stenberg, Antoine Stevens, Viacheslav Adamchuk, Hafedh Aïchi, Bernard G. Barthès, Harm M. Bartholomeus, Andreas D. Bayer, Martial Bernoux, Kristin Böttcher, Lukáš Brodský, Changwen Du, Adrian Chappell, Youcef Fouad, Vincent Genot, Cecile Gomez, Sabine Grunwald, Andreas Gubler, Carlos Guerrero, Carolyn B. Hedley, Maria Knadel, Héctor J. M. Morrás, Michele Nocita, Leonardo Ramirez-Lopez, Pierre Roudier, Eleodoro M. Rufasto Campos, Paul Sanborn, Victor M. Sellitto, Kenneth A. Sudduth, Barry G. Rawlins, Christian Walter, Leigh Ann Winowiecki, Suk Young Hong, and Wenjun Ji. A global spectral library to characterize the world's soil. *Earth-Science Reviews*, 155:198–230, 2016. doi: 10.1016/j.earscirev.2016.01.012.

[27] Jordi Inglada, Arthur Vincent, Marcela Arias, and Claire Marais-Sicre. Improved early crop type identification by joint use of high temporal resolution SAR and optical image time series. *Remote Sensing*, 8(5):362, 2016. doi: 10.3390/rs8050362.

[28] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, New York, 1998. Available online: https://www.wiley.com/en-us/Statistical+Learning+Theory-p-9780471030034 (accessed on May 2026).

[29] Farid Melgani and Lorenzo Bruzzone. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Transactions on Geoscience and Remote Sensing*, 42(8):1778–1790, 2004. doi: 10.1109/TGRS.2004.831865.

[30] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016. doi: 10.1145/2939672.2939785.

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

[32] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114, 2019. Available online: https://proceedings.mlr.press/v97/tan19a.html (accessed on May 2026).

[33] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018. doi: 10.1109/CVPR.2018.00745.

[34] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision*, pages 3–19, 2018. doi: 10.1007/978-3-030-01234-2_1.

[35] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016. doi: 10.1109/CVPR.2016.91.

[36] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLO (version 8.0.0). Computer software, 2023. Available online: https://github.com/ultralytics/ultralytics (accessed on May 2026).

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, 2015. doi: 10.1007/978-3-319-24574-4_28.

[38] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, volume 34, pages 12077–12090, 2021. Available online: https://proceedings.neurips.cc/paper_files/paper/2021/file/64f1f27bf1b4ec22924fd0acb550c235-Paper.pdf (accessed on May 2026).

[39] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022. doi: 10.1109/CVPR52688.2022.00135.

[40] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2961–2969, 2017. doi: 10.1109/ICCV.2017.322.

[41] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9 (8):1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

[42] Rai A. Schwalbert, Telmo Amado, Geomar Corassa, Luan P. Pott, P. V. Vara Prasad, and Ignacio A. Ciampitti. Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern brazil. *Agricultural and Forest Meteorology*, 284:107886, 2020. doi: 10.1016/j.agrformet.2019.107886.

[43] Olutobi Adeyemi, Ivan Grove, Sven Peets, and Tomas Norton. Dynamic neural network modelling of soil moisture content for predictive irrigation scheduling. *Sensors*, 18(10):3408, 2018. doi: 10.3390/s18103408.

[44] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018. Available online: https://arxiv.org/abs/1803.01271 (accessed on May 2026).

[45] Bryan Lim, Sercan O. Arik, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764, 2021. doi: 10.1016/j.ijforecast.2021.03.012.

[46] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. doi: 10.48550/arXiv.2312.00752.

[47] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. Available online: https://openreview.net/forum?id=YicbFdNTTy (accessed on May 2026).

[48] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. doi: 10.1109/ICCV48922.2021.00986.

[49] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10347–10357, 2021. Available online: https://proceedings.mlr.press/v139/touvron21a.html (accessed on May 2026).

[50] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. doi: 10.1109/CVPR52688.2022.01167.

[51] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. Available online: https://openreview.net/forum?id=nZeVKeeFYf9 (accessed on May 2026).

[52] Muhammad Awais, Ali Husain Salem Abdulla Alharthi, Amandeep Kumar, Hisham Cholakkal, and Rao Muhammad Anwer. AgroGPT: Efficient agricultural vision-language model with expert tuning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5687–5696, 2025. doi: 10.1109/WACV61041.2025.00555.

[53] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837, 2022. Available online: https://proceedings.neurips.cc/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html (accessed on May 2026).

[54] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022. Available online: https://openreview.net/forum?id=WE_vluYUL-X (accessed on May 2026).

[55] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474, 2020. Available online: https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html (accessed on May 2026).

[56] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009. doi: 10.1109/TNN.2008.2005605.

[57] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. Available online: https://arxiv.org/abs/1707.06347 (accessed on May 2026).

[58] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, pages 2672–2680, 2014. Available online: https://proceedings.neurips.cc/paper/2014/hash/f033ed80deb0234979a61f95710dbe25-Abstract.html (accessed on May 2026).

[59] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020. Available online: https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html (accessed on May 2026).

[60] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. doi: 10.1109/CVPR52688.2022.01042.

[61] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763, 2021. Available online: https://proceedings.mlr.press/v139/radford21a.html (accessed on May 2026).

[62] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. Available online: https://arxiv.org/abs/2303.05499 (accessed on May 2026).

[63] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. Available online: https://openreview.net/forum?id=a68SUt6zFt (accessed on May 2026).

[64] Ronja Güldenring and Lazaros Nalpantidis. Self-supervised contrastive learning on agricultural images. *Computers and Electronics in Agriculture*, 191:106510, 2021. doi: 10.1016/j.compag.2021.106510.

[65] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. Available online: https://openaccess.thecvf.com/content/ICCV2023/html/Kirillov_Segment_Anything_ICCV_2023_paper.html (accessed on May 2026).

[66] Inkyu Sa, Zetao Chen, Marija Popović, Raghav Khanna, Frank Liebisch, Juan Nieto, and Roland Siegwart. weedNet: Dense semantic weed classification using multispectral images and MAV for smart farming. *IEEE Robotics and Automation Letters*, 3(1):588–595, 2018. doi: 10.1109/LRA.2017.2774979.

[67] Andres Milioto, Philipp Lottes, and Cyrill Stachniss. Real-time semantic segmentation of crop and weed for precision agriculture robots leveraging background knowledge in CNNs. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 2229–2235, 2018. doi: 10.1109/ICRA.2018.8460962.

[68] Petteri Nevavuori, Nathaniel Narra, and Tarmo Lipping. Crop yield prediction with deep convolutional neural networks. *Computers and Electronics in Agriculture*, 163:104859, 2019. doi: 10.1016/j.compag.2019.104859.

[69] Jiaxuan You, Xiaocheng Li, Melvin Low, David Lobell, and Stefano Ermon. Deep gaussian process for crop yield prediction based on remote sensing data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, pages 4559–4565, 2017. Available online: https://ojs.aaai.org/index.php/AAAI/article/view/11172 (accessed on May 2026).

[70] Xiaoping Wu, Chi Zhan, Yu-Kun Lai, Ming-Ming Cheng, and Jufeng Yang. IP102: A large-scale benchmark dataset for insect pest recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8779–8788, 2019. doi: 10.1109/CVPR.2019.00899.

[71] Etienne David, Simon Madec, Pouria Sadeghi-Tehran, Helge Aasen, Bangyou Zheng, Shouyang Liu, Norbert Kirchgessner, Goro Ishikawa, Koichi Nagasawa, Minhajul A. Badhon, Curtis Pozniak, Benoit de Solan, Andreas Hund, Scott C. Chapman, Frédéric Baret, Ian Stavness, and Wei Guo. 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*, 2020:3521852, 2020. doi: 10.34133/2020/3521852.

[72] X. E. Pantazi, D. Moshou, T. Alexandridis, R. L. Whetton, and A. M. Mouazen. Wheat yield prediction using machine learning and advanced sensing techniques. *Computers and Electronics in Agriculture*, 121:57–65, 2016. doi: 10.1016/j.compag.2015.11.018.

[73] Afan Galih Salman, Bayu Kanigoro, and Yaya Heryadi. Weather forecasting using deep learning techniques. In *Proceedings of the 2015 International Conference on Advanced Computer Science and Information Systems*, pages 281–285, 2015. doi: 10.1109/ICACSIS.2015.7415154.

[74] Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619:533–538, 2023. doi: 10.1038/s41586-023-06185-3.

[75] Jiamei Liu, Fangle Chang, Jiahong Yang, Xinyi Jie, Caiyun Lu, Chao Wang, Lei Xie, Longhua Ma, and Hongye Su. Deep reinforcement learning for irrigation optimization: Advantages, opportunities, and challenges. *Agricultural Water Management*, 322:110030, 2025. ISSN 0378-3774. doi: https://doi.org/10.1016/j.agwat.2025.110030. URL https://www.sciencedirect.com/science/article/pii/S0378377425007449.

[76] Zeinab E. Mohamed, M. K. Afify, M. M. Badr, and O. A. Omar. IoT-driven smart irrigation system to improve water use efficiency. *Scientific Reports*, 16:2609, 2026. doi: 10.1038/s41598-025-33826-6.

[77] Tomislav Hengl, Jorge Mendes de Jesus, Gerard B. M. Heuvelink, Maria Ruiperez Gonzalez, Milan Kilibarda, Aleksandar Blagotić, Wei Shangguan, Marvin N. Wright, Xiaoyuan Geng,

Bernhard Bauer-Marschallinger, Mario Antonio Guevara, Rodrigo Vargas, Robert A. MacMillan, Niels H. Batjes, Johan G. B. Leenaars, Eloi Ribeiro, Ichsani Wheeler, Stephan Mantel, and Bas Kempen. SoilGrids250m: Global gridded soil information based on machine learning. *PLOS ONE*, 12(2):e0169748, 2017. doi: 10.1371/journal.pone.0169748.

[78] E.-C. Oerke. Crop losses to pests. *The Journal of Agricultural Science*, 144(1):31–43, 2006. doi: 10.1017/S0021859605005708.

[79] Junde Chen, Jinxiu Chen, Defu Zhang, Yuandong Sun, and Y. A. Nanehkaran. Using deep transfer learning for image-based plant disease identification. *Computers and Electronics in Agriculture*, 173:105393, 2020. doi: 10.1016/j.compag.2020.105393.

[80] Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. Learning skillful medium-range global weather forecasting. *Science*, 382:1416–1421, 2023. doi: 10.1126/science.adi2336.

[81] Anat Goldstein, Lior Fink, Amit Meitin, Shiran Bohadana, Oscar Lutenberg, and Gilad Ravid. Applying machine learning on sensor data for irrigation recommendations: Revealing the agronomist's tacit knowledge. *Precision Agriculture*, 19:421–444, 2018. doi: 10.1007/s11119-017-9527-4.

[82] Yang Yu, Kailiang Zhang, Li Yang, and Dongxing Zhang. Fruit detection for strawberry harvesting robot in non-structural environment based on Mask-RCNN. *Computers and Electronics in Agriculture*, 163:104846, 2019. doi: 10.1016/j.compag.2019.06.001.

[83] Ya Xiong, Cheng Peng, Lars Grimstad, Pål Johan From, and Volkan Isler. Development and field evaluation of a strawberry harvesting robot with a cable-driven gripper. *Computers and Electronics in Agriculture*, 157:392–402, 2019. doi: 10.1016/j.compag.2019.01.009.

[84] Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021. Available online: https://arxiv.org/abs/2104.13478 (accessed on May 2026).

[85] Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation- and translation-equivariant neural networks for 3d point clouds. *arXiv preprint arXiv:1802.08219*, 2018. Available online: https://arxiv.org/abs/1802.08219 (accessed on May 2026).

[86] Michael Grieves and John Vickers. Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems. In Franz-Josef Kahlen, Shannon Flumerfelt, and Anabela Alves, editors, *Transdisciplinary Perspectives on Complex Systems*, pages 85–113. Springer, Cham, 2017. doi: 10.1007/978-3-319-38756-7_4.

[87] German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019. doi: 10.1016/j.neunet.2019.01.012.

[88] Mike Davies, Narayan Srinivasa, Tsung-Han Lin, Gautham Chinya, Yongqiang Cao, Sri Harsha Choday, Georgios Dimou, Prasad Joshi, Nabil Imam, Shweta Jain, Yuyun Liao, Chit-Kwan Lin, Andrew Lines, Ruokun Liu, Deepak Mathaikutty, Steve McCoy, Arnab Paul, Jonathan Tse, Guruguhanathan Venkataramanan, Yi-Hsin Weng, Andreas Wild, Yoonseok Yang, and Hong Wang. Loihi: A neuromorphic manycore processor with on-chip learning. *IEEE Micro*, 38(1):82–99, 2018. doi: 10.1109/MM.2018.112130359.

[89] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. Available online: https://arxiv.org/abs/1503.02531 (accessed on May 2026).