International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 11 | Nov - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Intelligent Image Analysis Using Hybrid Deep Learning and Emerging Computational Techniques: A Comprehensive Review

V. Kanmani

Research Scholar, School of Computing Science

KPR College of Arts Science and Research, Coimbatore

Dr. S. Srividhya

Associate Professor & Head

Department of Information Technology

KPR College of Arts Science and Research, Coimbatore

Abstract

Image analysis has moved from hand-engineered descriptors to learned representations, and the current frontier is defined less by any single architecture than by deliberate combinations of them. This review examines hybrid deep learning, meaning models that couple two or more learning paradigms such as convolutional networks with transformers, recurrent units, classical classifiers, metaheuristic optimisers or symbolic reasoning, together with the emerging computational techniques that now shape how such models are trained and deployed. We organise the literature into a five-family taxonomy of hybrid designs and analyse eight emerging techniques: vision foundation models, diffusion-based generation, self-supervised learning, federated learning, edge and TinyML deployment, neuromorphic computing, quantum machine learning and explainable AI. We then map these methods onto six application domains, namely medical imaging, remote sensing, precision agriculture, autonomous systems, industrial inspection and visual security, and consolidate the benchmark datasets and evaluation metrics used to compare them. A comparative analysis of reported results shows a consistent pattern: hybrids that pair convolutional inductive bias with global attention tend to match or exceed pure designs at lower data and compute budgets, while foundation models shift the bottleneck from architecture design to adaptation and trust. We close with eight open challenges, covering data efficiency, robustness, interpretability, energy cost, privacy, fairness, standardised benchmarking and clinical or field validation, and propose a research agenda for each. The review is intended as a structured entry point for researchers and practitioners designing the next generation of intelligent image analysis systems.

Keywords: image analysis; hybrid deep learning; vision transformers; convolutional neural networks; foundation models; explainable AI; federated learning; edge computing

1. Introduction

Visual data now dominates the information produced by sensors, scanners, satellites and cameras, and turning it into reliable decisions is one of the central problems of applied artificial intelligence. Deep learning transformed this task after convolutional neural networks (CNNs) won the 2012 ImageNet challenge by a wide margin (Krizhevsky et al., 2017), and representation learning has since become the default foundation of computer vision (LeCun et al., 2015). Yet a decade of progress has also exposed the limits of any single model family. CNNs capture local texture efficiently but model long-range context poorly; vision transformers capture global context but are data-hungry (Dosovitskiy et al., 2021); and both remain opaque, energy-intensive and brittle under distribution shift.

1.1 Motivation

Hybrid deep learning responds to these limits by combining components whose strengths are complementary. A convolutional stem can supply locality and translation equivariance while a transformer encoder supplies global reasoning (Chen et al., 2021; Dai et al., 2021). A deep feature extractor can be paired with a margin-based classifier for small datasets, or tuned by a population-based optimiser when the hyperparameter space is large. In parallel, a set of emerging computational techniques, from large pretrained foundation models to federated, neuromorphic and quantum

computing, is changing what is feasible to train, where models can run and how far their outputs can be trusted. These two currents are now converging, which motivates a review that treats them together rather than separately.

1.2 Gap in existing reviews

Existing surveys are valuable but typically narrow in one of three ways. Some focus on a single architecture family, such as transformers in vision (Khan et al., 2022; Han et al., 2023). Others focus on a single domain, such as medical imaging (Litjens et al., 2017) or remote sensing (Zhu et al., 2017). A third group reviews one emerging technique in isolation, such as federated learning for health (Rieke et al., 2020). Few works offer a unified taxonomy of hybrid designs, connect them to the emerging techniques that determine their deployment, and compare them across domains with a common set of datasets and metrics.

1.3 Contributions

This review makes five contributions:

1. A **five-family taxonomy** of hybrid deep learning models for image analysis, with representative architectures, design rationale and trade-offs for each family.
2. A structured analysis of **eight emerging computational techniques** and how each interacts with hybrid model design.
3. A **cross-domain mapping** of methods to six application areas, with the dominant task types and representative studies.
4. A consolidated reference of **benchmark datasets, evaluation metrics and reported performance**, enabling like-for-like comparison.
5. A **research agenda** of eight open challenges, each paired with concrete directions.

1.4 Organisation

Section 2 traces the evolution of image analysis, and Section 3 summarises the foundational architectures on which hybrids are built. Section 4 presents the hybrid taxonomy and Section 5 the emerging techniques. Section 6 covers application domains, Section 7 datasets and metrics, and Section 8 a comparative analysis. Section 9 sets out open challenges and future directions, and Section 10 concludes.

2. Evolution of Image Analysis

Image analysis has passed through three broad eras, and each transition was driven by a change in where visual knowledge resides: first in the designer's features, then in learned weights, and now in pretrained, composable systems (Figure 1).

2.1 The hand-crafted era

Early pipelines separated feature extraction from decision making. Descriptors such as SIFT (Lowe, 2004) and HOG (Dalal & Triggs, 2005) encoded gradients and local structure, and a shallow learner, often a support vector machine (Cortes & Vapnik, 1995), produced the label. These systems were interpretable and data-efficient, but their accuracy plateaued because the features could not adapt to the task. LeNet-5 (LeCun et al., 1998) showed that features could instead be learned end to end, although hardware and data limited its reach at the time.

2.2 The deep CNN era

Two enablers, the ImageNet dataset (Deng et al., 2009) and GPU training, ended the plateau. AlexNet cut the top-5 error on ImageNet by roughly ten percentage points over the best hand-crafted entry (Krizhevsky et al., 2017). Successors deepened and regularised the design: VGG (Simonyan & Zisserman, 2015), residual networks (He et al., 2016), and compound-scaled EfficientNets (Tan & Le, 2019). Encoder-decoder networks such as U-Net (Ronneberger et al., 2015) extended the paradigm to dense prediction, and generative adversarial networks (Goodfellow et al., 2014) opened image synthesis.

2.3 The hybrid and foundation era

The transformer (Vaswani et al., 2017) replaced convolution with self-attention, and the Vision Transformer (ViT) applied it directly to image patches (Dosovitskiy et al., 2021). Because ViT lacks the locality prior of CNNs, it needs very large datasets, and this weakness triggered a wave of hybrids that reintroduce convolutional structure, including Swin (Liu et al., 2021), CoAtNet (Dai et al., 2021) and TransUNet (Chen et al., 2021). The same period saw vision-language pretraining (Radford et al., 2021), masked autoencoding (He et al., 2022), latent diffusion (Rombach et al., 2022) and promptable segmentation (Kirillov et al., 2023). Image analysis is increasingly built by adapting and combining large pretrained components rather than training a single network from scratch.

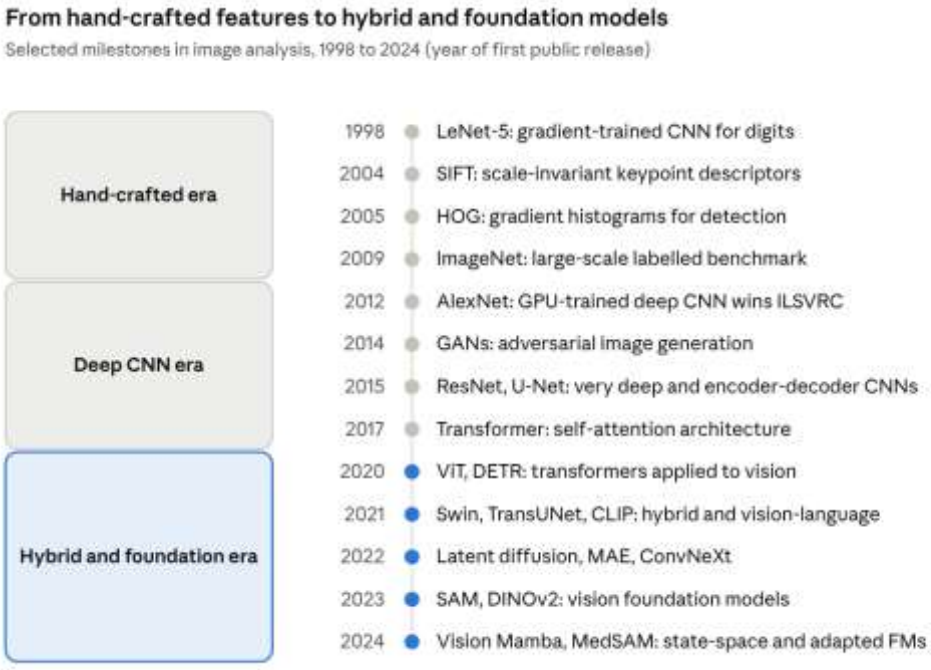


Figure 1. Milestones in image analysis (original figure)

Figure 1. Milestones in image analysis across three eras. Highlighted points mark the hybrid and foundation-model era reviewed in this paper.

3. Foundational Deep Architectures

Every hybrid is assembled from a small set of building blocks, and the choice of blocks follows from their inductive biases. This section summarises the four families most often combined; Table 1 contrasts them.

3.1 Convolutional neural networks

A convolutional layer applies a shared kernel across spatial positions, which yields translation equivariance and a strong locality prior. Stacking such layers builds a hierarchy from edges to parts to objects. Residual connections (He et al., 2016) made very deep stacks trainable, and depthwise-separable convolutions (Howard et al., 2017) made them cheap enough for mobile devices. ConvNeXt later showed that a carefully modernised CNN can rival transformers on ImageNet when trained with comparable recipes (Liu et al., 2022). For a feature map X, a convolution with kernel W computes:

$$Y(i,j) = \Sigma_m \Sigma_n W(m,n) \cdot X(i+m, j+n) + b \qquad (1)$$

3.2 Recurrent and sequence models

Recurrent networks, particularly long short-term memory units, model ordered dependencies. In image analysis they appear where an order exists: video frames, slices of a volumetric scan, satellite time series, or the words of a generated caption (Donahue et al., 2015; Vinyals et al., 2015). Their sequential computation limits parallelism, which is one reason attention has displaced them in many settings.

3.3 Transformers and attention

Self-attention lets each token weigh every other token, giving a global receptive field from the first layer. For queries Q, keys K and values V of dimension d, scaled dot-product attention is (Vaswani et al., 2017):

$$Attention(Q,K,V) = softmax(QK^{\mathrm{T}}/\sqrt{d}) \cdot V \qquad (2)$$

ViT splits an image into fixed-size patches and treats them as tokens (Dosovitskiy et al., 2021); DETR casts detection as set prediction with a transformer decoder (Carion et al., 2020). The cost of global attention grows quadratically with the number of tokens, a key motivation for windowed and hybrid designs. Structured state-space models offer a linear-time alternative, and Vision Mamba adapts them to images with bidirectional scanning (Zhu et al., 2024).

3.4 Generative models

Generative models learn the data distribution itself. Variational autoencoders optimise a likelihood bound (Kingma & Welling, 2014), GANs train a generator against a discriminator (Goodfellow et al., 2014), and denoising diffusion models learn to reverse a gradual noising process (Ho et al., 2020). In image analysis they serve three roles: synthetic data augmentation, image restoration and translation, and unsupervised anomaly detection.

Family	Core inductive bias	Main strengths	Main limitations	Typical role in a hybrid
CNN	Locality, weight sharing, translation equivariance	Data-efficient, fast, strong on texture and fine detail	Limited global context; receptive field grows slowly	Stem, local feature extractor, decoder
RNN / LSTM	Sequential order	Models temporal or slice-wise dependencies	Sequential computation; vanishing gradients over long spans	Temporal aggregator for video, volumes, captions
Transformer	Global pairwise interaction	Long-range context, scales with data, multimodal	Data-hungry; quadratic cost in token count	Global context encoder, cross-modal fusion
State-space (Mamba)	Selective linear recurrence	Linear-time long-sequence modelling	Young ecosystem; fewer validated vision results	Efficient replacement for attention blocks
Generative (VAE, GAN, diffusion)	Learned data distribution	Synthesis, restoration, anomaly scoring	Training instability (GAN); slow sampling (diffusion)	Data augmentation, pre-processing, prior

Table 1. Comparison of foundational deep architectures used as components in hybrid image analysis models.

4. Taxonomy of Hybrid Deep Learning Models

We define a hybrid model as one in which at least two components with different inductive biases or learning paradigms contribute to the same prediction, and are trained or tuned jointly. Under this definition, the literature falls into five families (Figure 2), distinguished by what is combined and why.

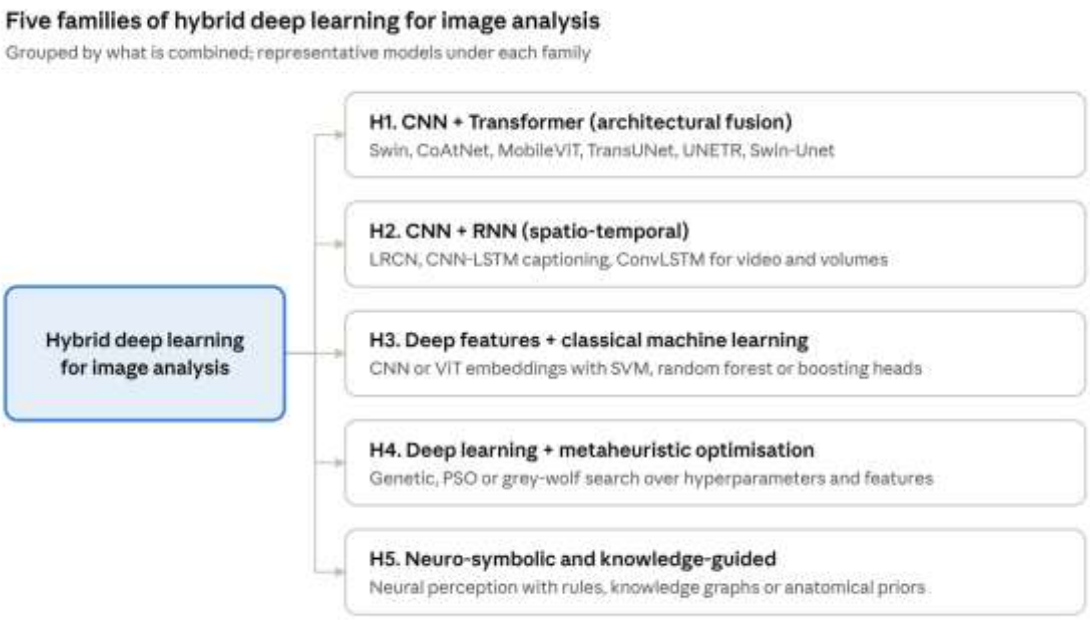


Figure 2. Taxonomy of hybrid deep learning models (original figure)

Proposed five-family taxonomy of hybrid deep learning for image analysis.

4.1 H1: CNN and transformer fusion

This is the most active family. Its rationale is complementary: convolution supplies locality and sample efficiency, while attention supplies global context. Three integration patterns recur. In **serial** hybrids, a CNN produces feature maps that are tokenised and passed to a transformer; TransUNet adds a U-Net decoder with skip connections for medical segmentation (Chen et al., 2021), and UNETR applies the idea to 3-D volumes (Hatamizadeh et al., 2022). In **interleaved** hybrids, convolution and attention alternate within stages, as in CoAtNet (Dai et al., 2021) and MobileViT, which targets mobile budgets (Mehta & Rastegari, 2022). In **attention-with-locality** designs, attention itself is restricted to shifted local windows, as in Swin (Liu et al., 2021) and its segmentation variant Swin-Unet (Cao et al., 2023). Figure 3 abstracts the common serial pattern.



Figure 3. Generic CNN-transformer hybrid architecture (original figure)

A generic serial CNN-transformer hybrid. The dashed skip path carries high-resolution convolutional features to the decoder, restoring detail lost during tokenisation.

4.2 H2: CNN and recurrent hybrids

When images arrive as sequences, a CNN encodes each frame and a recurrent unit aggregates them. Long-term recurrent convolutional networks applied this to activity recognition and video description (Donahue et al., 2015), and CNN-

LSTM pipelines became the standard for image captioning (Vinyals et al., 2015). Convolutional LSTMs preserve spatial structure inside the recurrence and remain common for precipitation nowcasting, satellite time series and slice-by-slice volumetric segmentation. Temporal transformers are steadily replacing the recurrent component, but CNN-RNN hybrids stay competitive where data and compute are limited.

4.3 H3: Deep features with classical machine learning

Here a pretrained network acts as a fixed or lightly tuned feature extractor, and a classical learner such as a support vector machine, random forest or gradient-boosted tree makes the decision. Replacing the softmax layer with a margin-based objective was shown early to improve generalisation on several benchmarks (Tang, 2013). The pattern is attractive in small-sample domains such as rare-disease imaging and plant pathology, because the classical head needs far fewer labelled examples and is easier to calibrate and audit.

4.4 H4: Deep learning with metaheuristic optimisation

Deep networks expose many discrete and continuous design choices. Population-based optimisers, including genetic algorithms, particle swarm optimisation and the grey wolf optimiser (Mirjalili et al., 2014), are used to select hyperparameters, prune features or search architectures. Neural architecture search generalises the idea: reinforcement-learning controllers (Zoph & Le, 2017) and regularised evolution (Real et al., 2019) have discovered image classifiers that match hand-designed ones. The main cost is compute, and reported gains are often modest once strong baselines are tuned with equal effort, a caveat many published studies omit.

4.5 H5: Neuro-symbolic and knowledge-guided hybrids

Neuro-symbolic systems couple neural perception with explicit knowledge: logical rules, scene graphs, ontologies or physical and anatomical constraints (Garcez & Lamb, 2023). In image analysis, examples include segmentation constrained by organ topology, visual question answering through executable programs, and remote-sensing classifiers regularised by land-cover rules. These hybrids trade some flexibility for interpretability and consistency, which makes them attractive in regulated settings.

Family	Hybridisation strategy	Representative models	Key advantage	Key limitation
H1 CNN + Transformer	Serial, interleaved or windowed attention	TransUNet, UNETR, CoAtNet, MobileViT, Swin, Swin-Unet	Local detail and global context at moderate data cost	Higher memory; design space is large
H2 CNN + RNN	Per-frame CNN encoder, recurrent aggregator	LRCN, CNN-LSTM captioners, ConvLSTM	Natural fit for video, volumes and time series	Sequential training; long-range forgetting
H3 Deep + classical ML	Deep embeddings feed SVM, random forest or boosting	CNN-SVM, ViT-feature ensembles	Works with small labelled sets; easier calibration	No end-to-end feature adaptation
H4 Deep + metaheuristics	Population search over hyperparameters, features or architectures	GA/PSO/GWO-tuned CNNs, NASNet, AmoebaNet	Automates design; can find compact models	High search cost; risk of unfair baselines
H5 Neuro-symbolic	Neural perception constrained or	Topology-aware segmentation,	Interpretable, consistent, data-efficient	Knowledge engineering

Family	Hybridisation strategy	Representative models	Key advantage	Key limitation
	explained by symbolic knowledge	program-based VQA		effort; brittle rules

Table 2. Summary of the five hybrid families.

5. Emerging Computational Techniques

Architecture is now only one lever. Eight techniques increasingly decide how image analysis models are pretrained, adapted, deployed and audited; Table 3 summarises how each one interacts with hybrid design.

5.1 Vision foundation models

Foundation models are pretrained once on very large, broad data and then adapted to many tasks. CLIP aligns images and text through contrastive learning on hundreds of millions of pairs and enables zero-shot classification (Radford et al., 2021). DINOv2 produces general-purpose visual features without labels (Oquab et al., 2024), and the Segment Anything Model (SAM) returns masks for arbitrary prompts after training on more than one billion masks (Kirillov et al., 2023). Domain adaptation is the new design problem: MedSAM fine-tunes SAM on over one million medical image-mask pairs and substantially improves on the general model in that domain (Ma et al., 2024). Many recent systems are therefore hybrids at a higher level, pairing a frozen foundation encoder with a small task-specific CNN, adapter or classical head.

5.2 Diffusion-based generation

Denoising diffusion probabilistic models generate images by learning to invert a fixed noising process (Ho et al., 2020). Latent diffusion runs this process in a compressed autoencoder space, cutting cost enough for high-resolution synthesis (Rombach et al., 2022), and diffusion transformers replace the U-Net backbone with a transformer (Peebles & Xie, 2023). For analysis rather than art, diffusion models are used to synthesise rare classes, to restore degraded or low-dose images, and to score anomalies by reconstruction error. Their main costs are slow iterative sampling and the risk that synthetic data leaks or amplifies biases in the training set.

5.3 Self-supervised learning

Self-supervised methods learn from unlabelled images by solving pretext tasks. Contrastive methods such as SimCLR pull together augmented views of the same image (Chen et al., 2020), while masked autoencoders reconstruct randomly hidden patches and scale efficiently with ViT backbones (He et al., 2022). Because annotation is the dominant cost in medicine, agriculture and industry, self-supervised pretraining on in-domain unlabelled data followed by light supervised fine-tuning has become a default recipe.

5.4 Federated learning

Federated learning trains a shared model across many data holders without moving raw data; each site sends model updates that are aggregated, for example by federated averaging (McMahan et al., 2017). It is especially relevant to medical imaging, where privacy rules and institutional policy often prevent data pooling (Kaissis et al., 2020; Rieke et al., 2020). The open problems are non-identically distributed data across sites, communication cost for large hybrid models, and residual privacy leakage through gradients, which is often addressed with secure aggregation or differential privacy.

5.5 Edge AI and TinyML

Many image analysis tasks, such as crop scouting, quality inspection and driver assistance, must run on cameras, drones or microcontrollers with tight power budgets. Efficiency comes from compact architectures such as MobileNets (Howard et al., 2017) and MobileViT (Mehta & Rastegari, 2022), from knowledge distillation into smaller students (Hinton et al., 2015), and from pruning and quantisation (Warden & Situnayake, 2019). Hybrid design matters here: a light convolutional stem with a few attention blocks often gives the best accuracy per milliwatt.

5.6 Neuromorphic computing

Neuromorphic hardware runs spiking neural networks that communicate with sparse, event-driven pulses, promising large energy savings for always-on vision (Roy et al., 2019). Chips such as Intel's Loihi implement on-chip learning with spiking neurons (Davies et al., 2018), and event cameras supply matching asynchronous input. Training spiking networks to deep-learning accuracy remains difficult, so current practice often converts a trained CNN into a spiking equivalent, itself a hybrid workflow (Schuman et al., 2022).

5.7 Quantum machine learning

Quantum machine learning explores whether quantum circuits can represent or process data more efficiently than classical models (Biamonte et al., 2017). Quantum convolutional neural networks were proposed as circuit analogues of CNNs (Cong et al., 2019), and hybrid quantum-classical networks typically insert a small variational circuit between classical layers. On current noisy hardware, demonstrations are limited to small, downsampled images, and no practical advantage over classical image models has yet been established. We therefore treat quantum image analysis as a long-horizon research direction rather than a deployable technique.

5.8 Explainable AI

As image models enter clinics, courts and factories, stakeholders need to know why a prediction was made. Post-hoc methods include Grad-CAM, which projects class-specific gradients onto convolutional feature maps (Selvaraju et al., 2017), and model-agnostic attributions such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017). Rudin (2019) argues that for high-stakes decisions, inherently interpretable models should be preferred over explaining black boxes. Hybrids can serve both views: attention maps and prototype layers make transformer-based hybrids partly self-explaining, while symbolic components in H5 models provide explicit reasons.

Technique	What it changes	Maturity for image analysis	Interaction with hybrid design	Representative works
Foundation models	Pretraining scale and transfer	High; widely adopted	Frozen encoder plus small task-specific hybrid head	Radford et al. (2021); Kirillov et al. (2023); Oquab et al. (2024)
Diffusion models	Generation, restoration, augmentation	High for synthesis; growing for analysis	Generative prior or data source for discriminative hybrids	Ho et al. (2020); Rombach et al. (2022)
Self-supervised learning	Label requirement	High	Pretrains CNN, ViT or hybrid backbones on unlabelled data	Chen et al. (2020); He et al. (2022)
Federated learning	Where data and training reside	Medium; pilots in healthcare	Communication cost scales with hybrid model size	McMahan et al. (2017); Rieke et al. (2020)
Edge AI and TinyML	Deployment hardware and budget	Medium to high	Favours light CNN stems with few attention blocks	Howard et al. (2017); Mehta & Rastegari (2022)
Neuromorphic computing	Energy and event-driven sensing	Low to medium	CNN-to-spiking conversion is a hybrid workflow	Davies et al. (2018); Roy et al. (2019)

Technique	What it changes	Maturity for image analysis	Interaction with hybrid design	Representative works
Quantum ML	Computational substrate	Low; small-scale demonstrations	Variational circuit inserted between classical layers	Biamonte et al. (2017); Cong et al. (2019)
Explainable AI	Trust, audit and debugging	Medium; post-hoc tools common	Attention, prototypes and symbolic parts aid explanation	Selvaraju et al. (2017); Rudin (2019)

Table 3. Emerging computational techniques and their relationship to hybrid image analysis models.

6. Application Domains

Hybrid models are not adopted uniformly: each domain favours the families that fit its data, constraints and risk profile. Table 4 maps the six domains reviewed here to their dominant tasks and model choices.

6.1 Medical imaging

Medical imaging is the largest consumer of hybrid designs because it combines small labelled datasets, high-resolution or volumetric inputs and a need for precise boundaries. CNN-transformer segmenters such as TransUNet, UNETR and Swin-Unet extend the U-Net lineage (Chen et al., 2021; Hatamizadeh et al., 2022; Cao et al., 2023), while the self-configuring nnU-Net remains a demanding CNN baseline that many hybrids fail to beat once evaluated fairly (Isensee et al., 2021). Dermatologist-level skin-lesion classification with a CNN (Esteva et al., 2017) and adapted foundation models such as MedSAM (Ma et al., 2024) illustrate the field's trajectory. Federated training and explainability are close to mandatory for clinical translation.

6.2 Remote sensing and Earth observation

Satellite and aerial imagery is multispectral, very large and often temporal (Zhu et al., 2017). CNN-transformer hybrids handle wide spatial context for land-cover segmentation, CNN-RNN and ConvLSTM hybrids model change over time, and knowledge-guided models encode geographic rules. Benchmarks such as EuroSAT, built from Sentinel-2 imagery, support land-use classification (Helber et al., 2019).

6.3 Precision agriculture

Image-based crop monitoring targets disease detection, weed mapping, yield estimation and fruit counting. Early work showed that deep CNNs could identify crop diseases from leaf images with high accuracy on the PlantVillage dataset (Hughes & Salathé, 2015; Mohanty et al., 2016), although accuracy falls sharply on field images with cluttered backgrounds. Lightweight hybrids deployed on drones and smartphones, combined with deep-feature plus classical-classifier pipelines for small farm-specific datasets, are the dominant response.

6.4 Autonomous systems

Self-driving vehicles and robots need real-time detection, segmentation and tracking under varying light and weather (Grigorescu et al., 2020). One-stage detectors such as YOLO (Redmon et al., 2016), two-stage detectors such as Faster R-CNN (Ren et al., 2015) and transformer detectors such as DETR (Carion et al., 2020) represent three design points on the speed-accuracy curve. Urban-scene datasets such as Cityscapes (Cordts et al., 2016) anchor evaluation, and spatio-temporal hybrids are used for trajectory prediction.

6.5 Industrial inspection

Manufacturing inspection must find rare, small defects on textured surfaces, usually with very few defective examples. Generative and reconstruction-based models learn normal appearance and flag deviations, while deep features with one-class classifiers form a strong H3 pattern. Edge deployment on production lines makes compact hybrids and quantisation essential.

6.6 Visual security and forensics

Security applications include face and person re-identification, surveillance analytics and detection of manipulated or synthetic images. The rise of diffusion-generated media makes forensic detection a moving target, and hybrids that combine frequency-domain cues with learned spatial features are a common defence. These applications carry significant privacy and fairness risks, so dataset governance and bias auditing are as important as accuracy.

Domain	Dominant tasks	Preferred hybrid families	Key constraints	Representative works
Medical imaging	Segmentation, classification, detection	H1, H3, H5; foundation models	Few labels, privacy, regulatory explainability	Esteva et al. (2017); Isensee et al. (2021); Ma et al. (2024)
Remote sensing	Land-cover segmentation, change detection	H1, H2, H5	Huge multispectral scenes, temporal series	Zhu et al. (2017); Helber et al. (2019)
Precision agriculture	Disease and weed detection, counting	H3, H4; edge-optimised H1	Field variability, low-cost hardware	Mohanty et al. (2016); Hughes & Salathé (2015)
Autonomous systems	Detection, segmentation, tracking	H1, H2	Real-time latency, safety, weather	Redmon et al. (2016); Carion et al. (2020); Grigorescu et al. (2020)
Industrial inspection	Anomaly and defect detection	H3; generative models	Rare defects, edge deployment	Ho et al. (2020); Howard et al. (2017)
Security and forensics	Re-identification, manipulation detection	H1, H3	Adversarial attackers, privacy, fairness	Rombach et al. (2022); Rudin (2019)

Table 4. Mapping of application domains to tasks, hybrid families and constraints.

7. Benchmark Datasets and Evaluation Metrics

Comparisons between hybrid models are only meaningful on shared datasets with consistent metrics. Table 5 lists the benchmarks most frequently used in the reviewed literature, and Table 6 defines the metrics.

7.1 Benchmark datasets

Dataset	Domain	Task	Scale and content	Source
ImageNet (ILSVRC-2012)	General objects	Classification	About 1.28 million training images, 50,000 validation images, 1,000 classes	Deng et al. (2009)
MS COCO	General scenes	Detection, instance segmentation, captioning	More than 200,000 labelled images, 80 object categories	Lin et al. (2014)

Dataset	Domain	Task	Scale and content	Source
Cityscapes	Urban driving	Semantic and instance segmentation	5,000 finely and 20,000 coarsely annotated frames from 50 cities	Cordts et al. (2016)
BraTS	Neuro-oncology MRI	Brain-tumour segmentation	Multimodal MRI (T1, T1c, T2, FLAIR) with expert tumour labels	Menze et al. (2015)
Synapse multi-organ CT	Abdominal CT	Multi-organ segmentation	30 contrast-enhanced CT scans, eight organs in common protocols	Chen et al. (2021)
MedMNIST v2	Biomedical	Classification (2-D and 3-D)	18 standardised datasets, about 708,000 2-D and 10,000 3-D images	Yang et al. (2023)
EuroSAT	Remote sensing	Land-use classification	27,000 Sentinel-2 patches, 10 classes, 13 spectral bands	Helber et al. (2019)
PlantVillage	Agriculture	Disease classification	About 54,000 leaf images across 14 crops and 38 crop-disease classes	Hughes & Salathé (2015)

Table 5. Benchmark datasets commonly used to evaluate hybrid image analysis models.

7.2 Evaluation metrics

Let TP, FP, TN and FN denote true positives, false positives, true negatives and false negatives. Classification is usually reported with accuracy, precision, recall and F1-score; segmentation with the Dice similarity coefficient and intersection over union; detection with mean average precision; and generation with the Fréchet inception distance. The two most common overlap measures for a predicted region A and ground-truth region B are:

$$Dice(A,B) = 2|A \cap B|/(|A| + |B|) \qquad (3)$$

$$IoU(A,B) = |A \cap B|/|A \cup B| \qquad (4)$$

Metric	Task	Definition	Interpretation
Accuracy	Classification	(TP + TN) / (TP + TN + FP + FN)	Misleading under class imbalance
Precision	Classification, detection	TP / (TP + FP)	Share of positive predictions that are correct
Recall (sensitivity)	Classification, detection	TP / (TP + FN)	Share of actual positives that are found
F1-score	Classification	2 × Precision × Recall / (Precision + Recall)	Balances precision and recall
AUC-ROC	Classification	Area under the true-positive vs false-positive rate curve	Threshold-independent ranking quality
Dice (DSC)	Segmentation	2 × overlap / (sum of region sizes)	Standard in medical segmentation
IoU / mIoU	Segmentation, detection	Overlap / union, averaged over classes	Standard in scene segmentation

Metric	Task	Definition	Interpretation
HD95	Segmentation	95th-percentile Hausdorff distance between boundaries	Penalises boundary outliers, in mm or pixels
mAP	Detection	Mean of per-class average precision over IoU thresholds	COCO standard averages IoU 0.50 to 0.95
FID	Generation	Distance between feature statistics of real and generated images	Lower is better
Params, FLOPs, latency	Efficiency	Model size, operations per image, inference time	Required for edge and fair comparison

Table 6. Evaluation metrics for image analysis tasks.

8. Comparative Analysis and Discussion

Across the reported evidence, hybrids that combine convolutional locality with attention deliver the best accuracy per parameter when training data is limited to ImageNet-1k, but the margin over a well-modernised CNN is small.

8.1 Classification efficiency

Figure 4 plots published ImageNet-1k results for models trained without extra data at 224-pixel resolution. At around 25 million parameters, CoAtNet-0 (81.6%) and Swin-T (81.3%) exceed the pure-transformer DeiT-S (79.8%) by 1.5 to 1.8 points. At the mobile end, MobileViT-S reaches 78.4% with 5.6 million parameters, ahead of EfficientNet-B0 at a similar size. However, ConvNeXt-T (82.1%) shows that a CNN redesigned with transformer-era training recipes closes the gap, which suggests that part of the reported advantage of early hybrids came from better training rather than architecture alone.

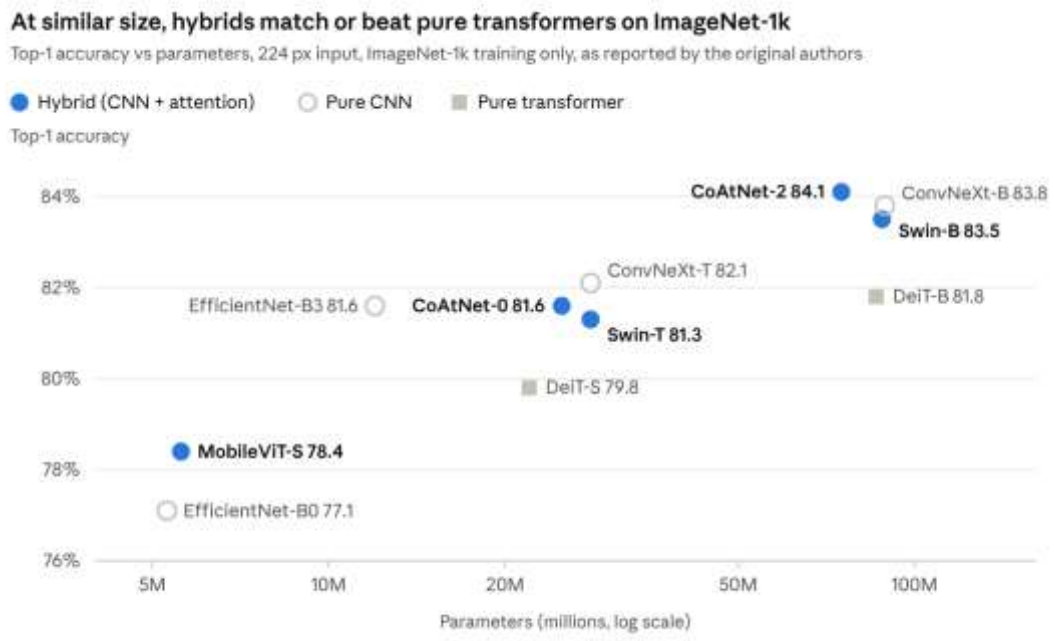


Figure 4. ImageNet-1k accuracy vs model size (data from the cited original papers)

Figure 4. ImageNet-1k top-1 accuracy against model size for representative CNN, transformer and hybrid models. Results are taken from the original papers and are not re-run under a single protocol.

8.2 Medical segmentation

On the Synapse multi-organ CT benchmark, the reported average Dice score rises from 76.85% for U-Net to 77.48% for TransUNet (Chen et al., 2021) and 79.13% for the fully windowed-attention Swin-Unet (Cao et al., 2023). These gains are real but modest, and independent re-evaluation indicates that a properly configured CNN such as nnU-Net remains highly competitive across many medical tasks (Isensee et al., 2021).

Model	Family	Synapse average DSC (%)	Source
Swin-Unet	H1 windowed attention	79.13	Cao et al. (2023)
TransUNet	H1 serial CNN-transformer	77.48	Chen et al. (2021)
U-Net	Pure CNN	76.85	As reported by Chen et al. (2021)

Table 7. Reported multi-organ segmentation performance on the Synapse CT dataset.

8.3 Synthesis of findings

1. **Inductive bias trades against data.** Pure transformers win only when pretraining data is very large; with modest data, hybrids or modern CNNs are the safer choice.
2. **Training recipes confound comparisons.** Augmentation, optimiser schedules and pretraining often explain as much variance as architecture, so reported gains below about one point should be treated cautiously.
3. **Foundation models shift the bottleneck.** When a frozen encoder such as DINOv2 or SAM is available, the main design question becomes adaptation, data curation and validation rather than backbone choice.
4. **Efficiency is under-reported.** Many studies omit FLOPs, latency or energy, which makes it hard to judge suitability for edge or clinical deployment.
5. **Metaheuristic and neuro-symbolic hybrids lack shared benchmarks.** Results are frequently reported on private or small datasets, which limits generalisable conclusions.

9. Open Challenges and Future Research Directions

The evidence reviewed points to eight unresolved challenges.

9.1 Data efficiency and annotation cost

Hybrids reduce but do not remove the need for labels. Promising directions include self-supervised pretraining on in-domain unlabelled archives, active learning that targets the most informative samples, and label-efficient adaptation of foundation models through adapters and prompt tuning.

9.2 Robustness and generalisation

Models that perform well on a benchmark often degrade under scanner changes, new geographies, weather or adversarial perturbation. Future work should report out-of-distribution and multi-site results as standard, and should study whether convolutional or attention components contribute more to robustness in a given hybrid.

9.3 Interpretability and trust

Post-hoc saliency maps can be unstable and do not always reflect the model's true reasoning. Inherently interpretable hybrids, such as prototype layers, concept bottlenecks and neuro-symbolic constraints, deserve more evaluation with domain experts rather than only visual inspection (Rudin, 2019).

9.4 Computational and energy cost

Training large hybrid and foundation models carries a significant energy and carbon cost. Directions include efficient attention and state-space blocks, distillation into compact students, hardware-aware architecture search and neuromorphic deployment for always-on sensing.

9.5 Privacy and data governance

Federated learning keeps raw data local, but gradient updates can still leak information. Combining federated hybrids with differential privacy, secure aggregation and synthetic-data sharing, and measuring the resulting accuracy cost honestly, remains an open research area (Kaissis et al., 2020).

9.6 Fairness and bias

Image datasets under-represent many populations, skin tones, crops and regions. Bias audits stratified by subgroup, balanced data collection and fairness-aware training objectives should become routine, particularly in medical and security applications.

9.7 Standardised and fair benchmarking

Many hybrid studies, particularly metaheuristic and domain-specific ones, compare against weak or untuned baselines on small private datasets. The field needs shared protocols that fix data splits, training budgets and reporting of variance, efficiency and statistical significance.

9.8 Clinical and field validation

Retrospective benchmark accuracy rarely translates directly into real-world benefit. Prospective studies, human-in-the-loop evaluation and post-deployment monitoring for drift are needed before hybrid systems can be trusted in clinics, farms and factories.

10. Conclusion

Intelligent image analysis is moving from a contest between architectures to the design of systems that combine complementary components. This review organised hybrid deep learning into five families: CNN-transformer fusion, CNN-recurrent models, deep features with classical learners, metaheuristic-optimised networks and neuro-symbolic designs. It then examined eight emerging techniques, from vision foundation models and diffusion to federated, neuromorphic and quantum computing, and explainable AI. Across six application domains, the evidence shows that hybrids pairing convolutional locality with attention achieve the best accuracy per parameter at moderate data scales, although carefully modernised CNNs remain strong baselines and many reported gains are small.

Three messages stand out for researchers and practitioners. First, the choice of components should follow the constraints of the problem, namely data volume, latency, privacy and the need for explanation, rather than architectural fashion. Second, foundation models are turning image analysis into a problem of adaptation, curation and validation, which raises the importance of rigorous evaluation. Third, progress now depends as much on fair benchmarking, efficiency reporting, fairness audits and prospective validation as on new architectures. Addressing the eight challenges identified here will determine whether hybrid image analysis systems move from impressive benchmarks to trustworthy, sustainable deployment.

References

Biamonte, J., Wittek, P., Pancotti, N., Rebentrost, P., Wiebe, N., & Lloyd, S. (2017). Quantum machine learning. *Nature, 549*(7671), 195–202. https://doi.org/10.1038/nature23474

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., & Wang, M. (2023). Swin-Unet: Unet-like pure transformer for medical image segmentation. In L. Karlinsky, T. Michaeli, & K. Nishino (Eds.), *Computer Vision – ECCV 2022 Workshops* (pp. 205–218). Springer. https://doi.org/10.1007/978-3-031-25066-8_9

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In *Computer Vision – ECCV 2020* (pp. 213–229). Springer. https://doi.org/10.1007/978-3-030-58452-8_13

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L., & Zhou, Y. (2021). *TransUNet: Transformers make strong encoders for medical image segmentation* (arXiv:2102.04306). arXiv. https://doi.org/10.48550/arXiv.2102.04306

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning* (Vol. 119, pp. 1597–1607). PMLR.

Cong, I., Choi, S., & Lukin, M. D. (2019). Quantum convolutional neural networks. *Nature Physics, 15*(12), 1273–1278. https://doi.org/10.1038/s41567-019-0648-8

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., & Schiele, B. (2016). The Cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3213–3223). https://doi.org/10.1109/CVPR.2016.350

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297. https://doi.org/10.1007/BF00994018

Dai, Z., Liu, H., Le, Q. V., & Tan, M. (2021). CoAtNet: Marrying convolution and attention for all data sizes. In *Advances in Neural Information Processing Systems* (Vol. 34, pp. 3965–3977). Curran Associates.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Vol. 1, pp. 886–893). https://doi.org/10.1109/CVPR.2005.177

Davies, M., Srinivasa, N., Lin, T.-H., Chinya, G., Cao, Y., Choday, S. H., Dimou, G., Joshi, P., Imam, N., Jain, S., Liao, Y., Lin, C.-K., Lines, A., Liu, R., Mathaikutty, D., McCoy, S., Paul, A., Tse, J., Venkataramanan, G., . . . Wang, H. (2018). Loihi: A neuromorphic manycore processor with on-chip learning. *IEEE Micro, 38*(1), 82–99. https://doi.org/10.1109/MM.2018.112130359

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255). https://doi.org/10.1109/CVPR.2009.5206848

Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2625–2634).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 542*(7639), 115–118. https://doi.org/10.1038/nature21056

Garcez, A. d'A., & Lamb, L. C. (2023). Neurosymbolic AI: The 3rd wave. *Artificial Intelligence Review, 56*(11), 12387–12406. https://doi.org/10.1007/s10462-023-10448-w

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems* (Vol. 27, pp. 2672–2680). Curran Associates.

Grigorescu, S., Trasnea, B., Cocias, T., & Macesanu, G. (2020). A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics, 37*(3), 362–386. https://doi.org/10.1002/rob.21918

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., Yang, Z., Zhang, Y., & Tao, D. (2023). A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(1), 87–110. https://doi.org/10.1109/TPAMI.2022.3152247

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H. R., & Xu, D. (2022). UNETR: Transformers for 3D medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 574–584). https://doi.org/10.1109/WACV51458.2022.00181

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 16000–16009). https://doi.org/10.1109/CVPR52688.2022.01553

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778). https://doi.org/10.1109/CVPR.2016.90

Helber, P., Bischke, B., Dengel, A., & Borth, D. (2019). EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 12*(7), 2217–2226. https://doi.org/10.1109/JSTARS.2019.2918242

Hinton, G., Vinyals, O., & Dean, J. (2015). *Distilling the knowledge in a neural network* (arXiv:1503.02531). arXiv. https://doi.org/10.48550/arXiv.1503.02531

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 6840–6851). Curran Associates.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). *MobileNets: Efficient convolutional neural networks for mobile vision applications* (arXiv:1704.04861). arXiv. https://doi.org/10.48550/arXiv.1704.04861

Hughes, D. P., & Salathé, M. (2015). *An open access repository of images on plant health to enable the development of mobile disease diagnostics* (arXiv:1511.08060). arXiv. https://doi.org/10.48550/arXiv.1511.08060

Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods, 18*(2), 203–211. https://doi.org/10.1038/s41592-020-01008-z

Kaissis, G. A., Makowski, M. R., Rückert, D., & Braren, R. F. (2020). Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence, 2*(6), 305–311. https://doi.org/10.1038/s42256-020-0186-1

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in vision: A survey. *ACM Computing Surveys, 54*(10s), Article 200. https://doi.org/10.1145/3505244

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. In *International Conference on Learning Representations*.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023). Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 4015–4026). https://doi.org/10.1109/ICCV51070.2023.00371

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM, 60*(6), 84–90. https://doi.org/10.1145/3065386

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278–2324. https://doi.org/10.1109/5.726791

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014* (pp. 740–755). Springer. https://doi.org/10.1007/978-3-319-10602-1_48

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis, 42*, 60–88. https://doi.org/10.1016/j.media.2017.07.005

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 10012–10022). https://doi.org/10.1109/ICCV48922.2021.00986

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 11976–11986). https://doi.org/10.1109/CVPR52688.2022.01167

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision, 60*(2), 91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 4765–4774). Curran Associates.

Ma, J., He, Y., Li, F., Han, L., You, C., & Wang, B. (2024). Segment anything in medical images. *Nature Communications, 15*, Article 654. https://doi.org/10.1038/s41467-024-44824-z

McMahan, B., Moore, E., Ramage, D., Hampson, S., & Agüera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics* (Vol. 54, pp. 1273–1282). PMLR.

Mehta, S., & Rastegari, M. (2022). MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer. In *International Conference on Learning Representations*.

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E., Weber, M.-A., Arbel, T., Avants, B. B., Ayache, N., Buendia, P., Collins, D. L., Cordier, N., . . . Van Leemput, K. (2015). The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging, 34*(10), 1993–2024. https://doi.org/10.1109/TMI.2014.2377694

Mirjalili, S., Mirjalili, S. M., & Lewis, A. (2014). Grey wolf optimizer. *Advances in Engineering Software, 69*, 46–61. https://doi.org/10.1016/j.advengsoft.2013.12.007

Mohanty, S. P., Hughes, D. P., & Salathé, M. (2016). Using deep learning for image-based plant disease detection. *Frontiers in Plant Science, 7*, Article 1419. https://doi.org/10.3389/fpls.2016.01419

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.-Y., Xu, H., Sharma, V., Li, S.-W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., . . . Bojanowski, P. (2024). DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*. https://doi.org/10.48550/arXiv.2304.07193

Peebles, W., & Xie, S. (2023). Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 4195–4205).

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning* (Vol. 139, pp. 8748–8763). PMLR.

Real, E., Aggarwal, A., Huang, Y., & Le, Q. V. (2019). Regularized evolution for image classifier architecture search. *Proceedings of the AAAI Conference on Artificial Intelligence, 33*(1), 4780–4789. https://doi.org/10.1609/aaai.v33i01.33014780

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 779–788). https://doi.org/10.1109/CVPR.2016.91

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems* (Vol. 28, pp. 91–99). Curran Associates.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778

Rieke, N., Hancox, J., Li, W., Milletarì, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R. M., Trask, A., Xu, D., Baust, M., & Cardoso, M. J. (2020). The future of digital health with federated learning. *npj Digital Medicine, 3*, Article 119. https://doi.org/10.1038/s41746-020-00323-1

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10684–10695). https://doi.org/10.1109/CVPR52688.2022.01042

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (pp. 234–241). Springer. https://doi.org/10.1007/978-3-319-24574-4_28

Roy, K., Jaiswal, A., & Panda, P. (2019). Towards spike-based machine intelligence with neuromorphic computing. *Nature, 575*(7784), 607–617. https://doi.org/10.1038/s41586-019-1677-2

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Schuman, C. D., Kulkarni, S. R., Parsa, M., Mitchell, J. P., Date, P., & Kay, B. (2022). Opportunities for neuromorphic computing algorithms and applications. *Nature Computational Science, 2*(1), 10–19. https://doi.org/10.1038/s43588-021-00184-y

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 618–626). https://doi.org/10.1109/ICCV.2017.74

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.

Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning* (Vol. 97, pp. 6105–6114). PMLR.

Tang, Y. (2013). *Deep learning using linear support vector machines* (arXiv:1306.0239). arXiv. https://doi.org/10.48550/arXiv.1306.0239

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning* (Vol. 139, pp. 10347–10357). PMLR.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 5998–6008). Curran Associates.

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3156–3164).

Warden, P., & Situnayake, D. (2019). *TinyML: Machine learning with TensorFlow Lite on Arduino and ultra-low-power microcontrollers*. O'Reilly Media.

Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., & Ni, B. (2023). MedMNIST v2: A large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data, 10*, Article 41. https://doi.org/10.1038/s41597-022-01721-8

Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., & Wang, X. (2024). Vision Mamba: Efficient visual representation learning with bidirectional state space model. In *Proceedings of the 41st International Conference on Machine Learning* (Vol. 235). PMLR.

Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., & Fraundorfer, F. (2017). Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine, 5*(4), 8–36. https://doi.org/10.1109/MGRS.2017.2762307

Zoph, B., & Le, Q. V. (2017). Neural architecture search with reinforcement learning. In *International Conference on Learning Representations*.