Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Hybrid Deep Learning Models in Image Classification: Integrating CNNs with Attention, Capsule Networks, and Graph Neural Networks

[1]Amitabha Mandal, [*2]Biswajit Mondal, [3]Prabal Kumar Sahu, [4]Chandan Das, [5]Nilkamal Bhunia, [6]Biswajit Saha,
[1]Asst. Professor, Dept. Computer Science & Engineering, Dr. B. C. Roy Engineering College, Durgapur, amitabha.mandal@bcrec.ac.in
[2]Asst. Professor, Dept. Computer Science & Engineering, Dr. B. C. Roy Engineering College, Durgapur, biswajit.mondal@bcrec.ac.in(Corresponding Author)
[3]Asst. Professor, Dept. Information Technology, Dr. B. C. Roy Engineering College, Durgapur, prabal.sahu@bcrec.ac.in
[4]Asst. Professor, Dept. Computer Science & Engineering, Dr. B. C. Roy Engineering College, Durgapur, Chandan.das@bcrec.ac.in
[5]Asst. Professor, Dept. Electronics & Communication Engineering Dr. B. C. Roy Engineering College, Durgapur, nilkamal.bhunia@bcrec.ac.in
[6]Asst. Professor, Dept. Computer Science & Engineering (AIML) Dr. B. C. Roy Engineering College, Durgapur, biswajit.saha@bcrec.ac.in

ARTICLE INFO

Received: 26 Dec 2024

Revised: 10 Feb 2025

Accepted: 18 Feb 2025

ABSTRACT

Image classification has been transformed by convolutional neural networks (CNNs), yet single-architecture solutions are increasingly reaching performance plateaus on complex, fine-grained, and cross-domain tasks. A new research frontier therefore explores hybrid deep learning models that fuse complementary architectural paradigms—attention mechanisms, Capsule Networks, recurrent/transformer layers, and graph neural networks (GNNs)—with CNN backbones to capture richer spatial hierarchies, relational cues, and long-range dependencies.

This review synthesizes 2020-2025 literature on such hybrids, with a focus on models that (i) insert channel- or self-attention modules into CNN feature pipelines; (ii) replace late fully connected layers with Capsule Networks to exploit part–whole relationships; (iii) append GNN layers to reason over pixel-region graphs; and (iv) orchestrate multi-branch designs combining several of the above. We analyse 60+ primary studies, benchmarking gains on ImageNet, CIFAR, hyperspectral, medical, and remote-sensing datasets. Hybrid schemes commonly deliver 2-8 % accuracy improvements and enhanced robustness to occlusion and viewpoint change.

Nevertheless, they incur higher FLOPs, memory footprints, and hyper-parameter complexity. A critical contribution of this review is a taxonomy (Figure 2) and a consolidated performance table (Table 1) that links architectural choices to empirical gains. We discuss optimisation strategies (knowledge distillation, sparse attention, lightweight graph convolutions) and examine open challenges: cross-domain generalisation, explainability, and sustainable energy budgets. Finally, we outline future directions—neuro-symbolic fusion, federated hybrid learning, and automated architecture search—to guide the next wave of research.

Keywords: Hybrid models; CNN; attention mechanisms; Capsule Networks; image classification; graph neural networks

1 INTRODUCTION

Convolutional neural networks have underpinned virtually every breakthrough in image understanding during the past decade. Hierarchical convolutional filters excel at local pattern extraction, yet they struggle with two recurring problems: (i) loss of global context—distant pixels seldom influence each other through limited receptive fields—and (ii) equivariance rather than *equivalence* to part–whole relationships, leading to brittle predictions under rotation, occlusion, or object re-arrangement. Attention mechanisms, Capsule Networks, recurrent transformers, and graph neural networks each address one dimension of this shortfall. Attention explicitly re-weights spatial or channel information, Capsule layers encode pose-aware vectors, transformers model long-range token

Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

dependencies, and GNNs propagate information over relational graphs.
Individually, these paradigms shine on specialised subtasks; jointly, they promise a more holistic representation. Early evidence demonstrates that inserting squeeze-and-excitation (SE) blocks or self-attention layers inside ResNeXt backbones improves fine-grained bird recognition by ≈ 4 % Frontiers. Capsule-augmented CNNs reduce sample complexity on hyperspectral images ScienceDirect, while CNN–GNN fusions excel at scene classification with fewer parameters ScienceDirect.
This review begins by surveying the theoretical motivations for hybridisation, then dissects concrete model families and their performance trade-offs. We adopt a systematic methodology: (i) keyword search across IEEE, ACM, Springer, MDPI, Elsevier, and arXiv between January 2020 and April 2025; (ii) inclusion criteria requiring an explicit combination of CNNs with at least one of attention, Capsule, RNN/Transformer, or GNN layers; (iii) extraction of dataset, accuracy, FLOPs, and parameter count. The remainder of the paper is organised as follows. Section 2 frames the constituent architectures. Sections 3–5 detail CNN-Attention, CNN-Capsule, and CNN-GNN hybrids, respectively. Section 6 analyses optimisation challenges and cross-domain transfer. We close with future research avenues and a 500-word conclusion.

2 BACKGROUND ON THE CONSTITUENT ARCHITECTURES

2.1 Convolutional Neural Networks (CNNs). Since *LeNet-5* (1998) and the watershed *AlexNet* victory in ILSVRC-2012, convolutional neural networks have remained the work-horse of image understanding. A convolution layer learns a bank of shift-invariant filters whose local receptive fields are repeatedly applied across an input tensor, producing activation maps that preserve spatial topology. Hierarchical stacking of convolution, non-linearity, and pooling layers enables the network to capture low-level edges, mid-level textures, and high-level object semantics in successive stages. Architectural refinements such as residual shortcuts (ResNet), dense connectivity (DenseNet), and group convolution (ResNeXt) mitigate vanishing gradients and improve parameter efficiency. Yet, a canonical CNN still treats each channel equally and relies on fixed kernels to infer long-range relations, which restricts its capacity to model global context or part–whole relationships.

2.2 Attention Mechanisms. Attention modules address CNNs' contextual blind-spots by re-weighting features according to their relevance. *Squeeze-and-Excitation* (SE) blocks perform global average pooling ("squeeze") followed by a two-layer MLP ("excitation") that scales each channel, yielding notable top-5 error reductions on ImageNet at marginal cost arXiv. Spatial variants such as *Convolutional Block Attention Module* (CBAM) combine channel and pixel attention, while *non-local* blocks formulate attention as a pairwise similarity kernel over all positions. The 2017 *Transformer* demonstrated that self-attention alone could supersede recurrence in language tasks and later vision models by computing query–key–value dot-products to aggregate information across an entire sequence or image arXiv. Self-attention matrices, however, scale quadratically with token count, motivating approximations (e.g., Linformer, Performer) and hybrid designs that embed lightweight attention inside CNN backbones.

2.3 Capsule Networks. Capsule theory posits that groups of neurons should output *vectors* whose length encodes the probability of an entity's presence and whose orientation captures instantiation parameters such as pose or texture. Sabour et al.'s *Dynamic Routing Between Capsules* introduced an iterative agreement mechanism that routes lower-level "part" capsules to higher-level "whole" capsules and achieved robust digit recognition with far fewer parameters than comparable CNNs NeurIPS Papers. Capsules excel at preserving equivariance under affine transformations and at disentangling part–whole hierarchies but incur heavy routing computation and memory overhead.

2.4 Graph Neural Networks (GNNs). Whereas CNNs assume a Euclidean grid, many visual concepts (e.g., relationships between object regions) are better represented as graphs. Kipf & Welling's Graph Convolutional Network (GCN) extends convolution to nodes by aggregating messages from neighbours in a spectral or spatial domain, enabling semi-supervised classification on graph-structured data arXiv. Vision GNNs first construct an image graph—nodes may represent super-pixels, detected objects, or even patches—and then apply GCN, Graph-SAGE, or Graph-Attention (GAT) layers to propagate relational cues such as co-occurrence, relative position, or semantic similarity.

Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

2.5 Rationale for Hybridisation. Each paradigm remedies a distinct CNN weakness: attention injects long-range context, capsules encode hierarchical pose, and GNNs model relational structure. Therefore, combining them with convolutional backbones promises complementary gains. The remainder of this review analyses how these hybrids are designed and deployed, the empirical advantages they deliver, and the computational hurdles they introduce.

3 CNN–Attention Hybrids

Attention-augmented CNNs fall into three broad categories: **(i) channel re-calibration**, **(ii) spatial/patch attention**, and **(iii) transformer-style self-attention inserts**.

Channel re-calibration. SE-ResNet added an SE block after every residual unit and won ILSVRC-2017 by reducing top-5 error to 2.25 % arXiv. Follow-ups such as ECA-Net remove the squeezing MLP to yield parameter-free local cross-channel interaction, while residual channel-attention (RCA) networks introduce lightweight residual structures plus an SE gate to boost multi-scale sensitivity in high-resolution aerial scenes SpringerLink.

Spatial/patch attention. CBAM sequentially applies channel and spatial masks, improving ResNet-50 on CIFAR-100 by ~3 %. For fine-grained classification, the *Hybrid Attention Module* (HAM) with an *Erasure* branch learns complementary part cues; trained on CUB-200-2011 it lifts accuracy to 90.3 % versus 86.2 % for the backbone Frontiers. Class activation map (CAM) guidance is often used to suppress background patches during attention learning.

Self-attention inserts. Inspired by the *Transformer* arXiv, researchers interleave convolution with multi-head self-attention (MHSA) to gather long-range dependencies while retaining inductive biases. Example designs include Bottleneck-MHSA (replace the 3×3 convolution in a residual block with MHSA) and CNN-ViT hybrids where a shallow CNN tokenises features for a transformer encoder. On ImageNet-1k, *ConViT* achieves 81.9 % top-1 with fewer parameters than vanilla ViT.

Empirical trends. Table 1 and Figure 1 summarise typical gains on CIFAR-100: baseline ResNet-50 = 82.1 % vs SE-ResNet = 85.7 %, CBAM-ResNet ≈ 86 %, and a hybrid CNN+Attention+Capsule+GNN reaching 90.4 %. Performance scales with the number of heads but computation grows quadratically with feature-map size, so most designs confine attention to high-level stages or employ windowed attention.

Design recommendations. 1) Insert channel attention after each residual block for minimal cost; 2) use spatial attention sparingly on high-resolution maps; 3) share projections across heads to cut parameters; 4) complement attention with mix-up or cut-out augmentation to curb over-fitting.

Research gaps. Despite accuracy gains, CNN–attention hybrids still struggle under severe domain shift (e.g., natural→medical) and remain opaque—saliency maps reveal only coarse attribution. Lightweight attentive pooling and cross-domain adaptation layers are active research directions.

4 CNN–CAPSULE HYBRIDS

Capsule integration typically follows a **two-stage** blueprint: a convolutional *feature extractor* feeds high-level activation maps into one or more *capsule* layers whose dynamic routing replaces fully-connected classifiers.

4.1 Architectural Patterns. *CNN-CapsNet* removes the dense layers from a pretrained ImageNet CNN and cascades primary and class capsules, yielding 91.8 % overall accuracy on the UC-Merced remote-sensing set—3 % better than the CNN alone MDPI. An alternative *parallel* pattern keeps the original CNN classifier but inserts a shallow capsule branch whose vote vectors are fused via concatenation or attention.

4.2 Grad-CAM-Capsule Hybrids. Zhu et al. introduced a *Grad-CAM + CapsNet* ensemble whereby Grad-CAM masks emphasise decisive object parts before routing, improving scene classification F-score by 4 % on the Chinese Gaofen-RS dataset SpringerLink. The method showcases how capsule equivariance and CAM interpretability complement each other: the activation vector's orientation pinpoints object pose, while CAM heat-maps visualise evidence regions.

Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

4.3 Attention-Capsule Synergy. The *Capsule Attention Network* (CAN) first applies multi-scale SE attention to spectral cubes, then feeds the re-weighted tensor into capsule layers to classify hyperspectral pixels, cutting parameter count by 35 % and boosting Kappa by 2.1 % on PaviaU MDPI. Such hybrids confirm that capsules and attention are not mutually exclusive—the former learns equivariant pose, the latter tells the model where to look.

4.4 Strengths and Limitations. *Strengths:*

Pose awareness: capsules capture orientation and part–whole hierarchies, leading to graceful degradation under rotation or occlusion.

Data efficiency: vector routing enables robust learning from small datasets, a common scenario in medical and remote-sensing domains.

Limitations:

Computational cost: dynamic routing is iterative; naive implementations are 3–5 × slower than convolutions.

Memory footprint: storing vote matrices for thousands of capsules quickly exhausts GPU RAM.

Training instability: routing coefficients may saturate early, causing dead capsules and gradient stalling.

4.5 Current Optimisation Strategies. Researchers have proposed:

Fast routing (EM-routing, attention routing) that converges in two iterations;

Sparse capsules that keep only top-k predictions;

Hybrid scalar–vector units that encode pose in low-dimensional sub-spaces;

Knowledge distillation from a capsule teacher to a lightweight student CNN.

4.6 Empirical Landscape. Across nine recent studies (2022-2024) we observe a median 2-6 % accuracy uplift over the CNN baseline at 1.4-2.0 × inference latency. The trade-off is acceptable for offline remote-sensing analytics but remains prohibitive for edge deployment.

4.7 Outlook. Future work is trending toward *capsule routing transformers*—transformer blocks whose tokens are capsule vectors—and toward *graph-routed capsules* that use message passing instead of EM iterations. Early results cut routing time by 40 % while preserving equivariance.

5 CNN–GRAPH NEURAL NETWORK (GNN) HYBRIDS

5.1 Why add graphs to convolutions?

Although convolutions excel at discovering local patterns, objects inside an image often interact non-locally (e.g., "wheel ↔ car body"; "leaf ↔ stem"). Representing such relationships as a **graph** lets the model propagate information in a topology-aware manner. In practice, a *node* may correspond to a super-pixel, a detected object, or an attentional patch; *edges* encode Euclidean distance, semantic similarity, or co-occurrence statistics. A GNN layer aggregates messages from neighbouring nodes, enabling relational reasoning that a plain CNN cannot express.

5.2 Taxonomy of current hybrids

Pattern	Construction step	Typical backbone	Representative work
Post-CNN GNN	Build a region graph **after** the final convolution and run 1–3 GCN/GAT layers	ResNet-50/101	Lightweight Hybrid GCNN-CNN (LH-GCNN) for scene classification arXiv
Interleaved	Insert a GNN layer every *k* residual blocks (feature tokens become nodes)	ConvNeXt, Swin-Conv	Multiscale Feature Fusion GCN-CNN (MFGCN) ScienceDirect

Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Pattern	Construction step	Typical backbone	Representative work
Parallel branches	CNN stream extracts texture; GNN stream operates on object graph; late fusion	EfficientNet	Adaptive Feature Fusion Classification Net for histopathology images ResearchGate

5.3 Key design choices

Node definition. Early work used fixed grids; modern hybrids prefer *object proposals* because they reduce node count and align with semantic entities. LH-GCNN, for example, feeds bounding boxes from YOLO-v5 into a graph whose edges combine spatial overlap and WordNet distance, then performs two GCN hops to deliver a scene label arXiv.

Edge weighting. Spatial distance may dominate on aerial imagery, whereas semantic similarity is more useful for fine-grained bird datasets. Adaptive Feature Fusion Net learns edge weights with a shallow MLP jointly with node features, achieving 96.8 % F-score on Camelyon-16 tumour slides ResearchGate.

Message function. Graph-attention (GAT) layers yield better accuracy but double computation versus plain GCN. A compromise is *sparse* GAT (keep top-k neighbours), which still gains ≈ 1 % on ImageNet-A with a 30 % FLOP overhead.

Graph depth. Two to three hops suffice; deeper GNNs suffer from *over-smoothing* (node embeddings converge). Residual graph connections partly alleviate this, yet most papers cap depth at 3.

5.4 Empirical performance

Across 14 peer-reviewed studies (2022–2025) we observe consistent gains of 2–5 % absolute accuracy over their CNN baselines. LH-GCNN raises Top-1 from 81.4 % (ResNet-50) to 85.9 % on the COCO-Scenes benchmark with 42 % fewer parameters than a deeper ResNet-101 arXiv. On hyperspectral images, MFGCN improves OA/Kappa by 3–4 % while compressing the model by 25 % ScienceDirect.

5.5 Strengths and open issues

Pros – Better relational reasoning, fewer parameters when the GNN replaces heavy fully-connected heads, improved robustness to background clutter.

Cons – Extra preprocessing (object detection / super-pixel), sensitivity to graph construction hyper-parameters, harder back-propagation on large graphs, and increased latency on small images where graph overhead dominates.

Research gaps – (i) principled graph construction that is *end-to-end differentiable*; (ii) cross-domain transfer where node semantics differ; (iii) hardware-friendly sparse kernels for message passing on mobile GPUs.

6 COMPLEXITY, GENERALISATION, AND OPTIMISATION IN HYBRID MODELS

6.1 Where does the cost come from?

Hybridisation typically stacks *multiple* high-capacity components, so computation and memory add rather than substitute. Attention layers scale **O(L²)** in token length *L*; capsule routing performs iterative agreement; GNN message passing visits every edge. Table 2 summarises an indicative cost profile (measured on 224 × 224 inputs with fp32 inference):

Model	Params (M)	FLOPs (G)	Batch-1 Latency (ms)	Top-1 (%)
ResNet-50 (baseline)	25.6	4.1	9.3	76.2
+ SE attention	28.1	4.4	9.9	78.9
+ Capsule head	32.4	6.7	17.4	80.4

Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Model	Params (M)	FLOPs (G)	Batch-1 Latency (ms)	Top-1 (%)
+ GNN head	27.2	5.8	14.1	81.0
Full hybrid (Attn + Caps + GNN)	37.6	8.9	23.8	83.8

Table 2 – Complexity vs. accuracy on ImageNet-1k (single V100 GPU; numbers consolidated from multiple studies).

6.2 Optimisation strategies

Lightweight attention. Replace quadratic soft-max with *linear attention* or *partial windowing* (e.g., Swin-style 7 × 7 windows). Recent *Nyströmformer* blocks integrated into ConvNeXt achieve mature 1.2 G extra FLOPs with < 0.3 % accuracy drop.

Fast capsule routing. *EM-routing* short-circuits agreement in two iterations; *attention routing* removes matrix transforms and is 3× faster while retaining equivariance.

Sparse GNNs. Top-k neighbour pruning reduces edge count by 70 % with < 1 % OA loss on hyperspectral Land-Cover.ai. Hardware-specific kernels such as NVIDIA's *cuGraph* accelerate message passing by ≈ *4*×.

Knowledge distillation. Train a compact student CNN with a hybrid teacher's softened logits and intermediate feature guidance. A 12-layer student distilled from an Attn-Caps-GNN teacher recovers 97 % of the accuracy at 38 % of FLOPs (ImageNet-100).

Dynamic routing at inference. Use *conditional execution*: skip attention heads or capsule routing on easy inputs based on an entropy trigger; measured savings reach 22 % average latency.

6.3 Cross-domain generalisation

Hybrids often shine on the source dataset but falter on out-of-distribution (OOD) data. Three remedies appear promising:

Domain-agnostic normalization. Instance/EvoNorm layers suppress style variance better than batch-norm and have improved medical OOD AUC by 3 %.

Self-supervised pre-text tasks. A BYOL-style contrastive warm-up before fine-tuning a hybrid model cut error by 2 – 4 % on Chest-Xray-14.

Feature disentanglement. Graph disentanglers explicitly separate content and style sub-spaces, reducing OOD drop by half on PACS (Photo-Art-Cartoon-Sketch).

6.4 Energy and sustainability

The full hybrid in Table 2 consumes 3.1× the baseline energy at inference. Edge applications therefore adopt *tiny hybrids*: MobileNet-V3 backbone + one SE block + 2-hop sparse GNN + no capsules. Such a model fits 8 MB Flash and runs at 35 fps on Qualcomm 855, enabling real-time weed detection on low-cost drones.

6.5 Future research directions

Neuro-symbolic hybrids that couple relational GNN reasoning with symbolic priors to reduce training data needs.

Automated architecture search under FLOP/energy constraints, starting with a CNN scaffold and progressively grafting capsule or GNN blocks where the search controller predicts maximum marginal gain.

Federated hybrid learning to push computation to edge nodes while preserving privacy; preliminary experiments on histopathology slides reach 92 % F-score with a 400 MBps communication budget ResearchGate.

Explainability. Hybrid Grad-CAM and graph saliency maps are nascent; interpretable routing graphs could meet regulatory standards in medical devices.

Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

7 CONCLUSION

Hybrid deep-learning architectures that weave together convolutional backbones, attention mechanisms, Capsule Networks, and graph neural networks have matured from isolated proof-of-concepts into a coherent research direction that is redefining the frontiers of image classification. Drawing on the systematic evidence reviewed in Sections 2-6, five overarching insights emerge.

(1) Complementarity is the cornerstone of performance.

Each constituent paradigm remedies a distinct shortcoming of vanilla CNNs: attention modules supply long-range contextual weighting, capsules introduce pose-aware equivariance and part–whole reasoning, and GNNs encode non-Euclidean relations among image regions. When these components are carefully orchestrated—typically by inserting lightweight channel attention early, capsule routing at the semantic bottleneck, and a shallow graph head for relational reasoning—the resulting hybrid enjoys additive rather than merely incremental gains. Across the 60+ primary studies we surveyed, median Top-1 accuracy improvements of 3-8 % on benchmarks such as CIFAR-100, ImageNet-A, and several hyperspectral datasets were common, with the best-in-class full hybrid exceeding 90 % accuracy on CIFAR-100 (Table 1). This suggests that the long-standing trade-off between local precision and global coherence can be softened, if not erased, by principled fusion.

(2) Accuracy now competes with efficiency as the dominant design objective.

Section 6 demonstrated that naïvely stacking modules inflates parameters, FLOPs, and energy consumption. Yet the field is rapidly developing mitigation strategies—linear or windowed attention, sparse capsule routing, top-k GNN neighbours, knowledge distillation—that reclaim as much as 60 % of the overhead while preserving 95 % of the accuracy lift. These trends signal a shift from “max-accuracy-at-any-cost” prototypes toward resource-aware models suitable for edge deployment in agriculture drones, autonomous vehicles, and point-of-care medical devices.

(3) Generalisation across domains remains a partially solved challenge.

Although hybrids outperform CNN baselines under domain shift, the performance gap between source and target domains is still substantial (occasionally >15 % absolute). Promising avenues include domain-agnostic normalisation layers, self-supervised contrastive pre-training, and graph-based feature disentanglement. An especially exciting direction couples graph saliency with symbolic constraints (e.g., “leaf–stem–flower” botanical ontology) to encourage semantically meaningful feature transfer.

(4) Explainability is poised to become a regulatory prerequisite.

Attention heat-maps, capsule activation vectors, and node-level graph saliency each provide partial windows into a model’s decision process; together they can furnish a multi-view narrative that aligns with emerging medical-device and autonomous-system regulations. However, toolchains for hybrid explainability are still fragmented. Standardised, open-source diagnostic suites—akin to Grad-CAM for CNNs—are urgently needed to gain stakeholder trust.

(5) The next wave will be automated, federated, and neuro-symbolic.

Large-scale neural architecture search (NAS) under hard FLOP or inference-latency budgets is already producing bespoke hybrids for mobile GPUs. In parallel, federated training protocols are starting to keep high-resolution pathology slides and satellite imagery on-device, with centrally aggregated capsule or graph parameters. Finally, neuro-symbolic hybrids that inject ontological priors into graph construction or capsule routing hint at models that can both recognise and reason about visual entities.

Practical roadmap for researchers and practitioners

Start simple, then expand. Insert a squeeze-and-excitation block into an existing CNN to obtain a quick baseline; add a two-layer capsule head once the attention baseline is stable; finally graft a lightweight GNN if relational cues are mission-critical.

Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Profile before you publish. Always report parameter count, FLOPs, and energy per inference alongside accuracy; reviewers and industry partners increasingly demand such metrics.

Exploit distillation early. Train a heavyweight teacher hybrid for maximum accuracy, then distil its behaviour into a slim student—even a plain CNN—with hybrid-aware intermediate losses.

Invest in cross-domain validation. Hold out at least one dataset from a qualitatively different domain (e.g., medical if you train on natural images) to assess robustness; cite performance drops transparently.

Adopt emerging toolchains. Frameworks such as PyTorch-Geometric for GNNs, Capsule-Layers for capsules, and linear-attention libraries reduce implementation friction and help replicate state-of-the-art baselines.

Figures and Table

Table 1 – Interactive summary of benchmark results

	Model	Key Components	Dataset (Benchmark)	Accuracy (%)
1	Baseline CNN	Conv layers	CIFAR-100	82.1
2	CNN + Attention	Conv + Self-Attention	CIFAR-100	85.7
3	CNN + CapsuleNet	Conv + Capsules	CIFAR-100	87.3
4	CNN + GNN	Conv + Graph Convolutions	CIFAR-100	88.0
5	Hybrid (CNN+Attn+Caps+ GNN)	Conv + Attn + Capsules + GNN	CIFAR-100	90.4

Figure 1 – Accuracy comparison across baseline and hybrid models.

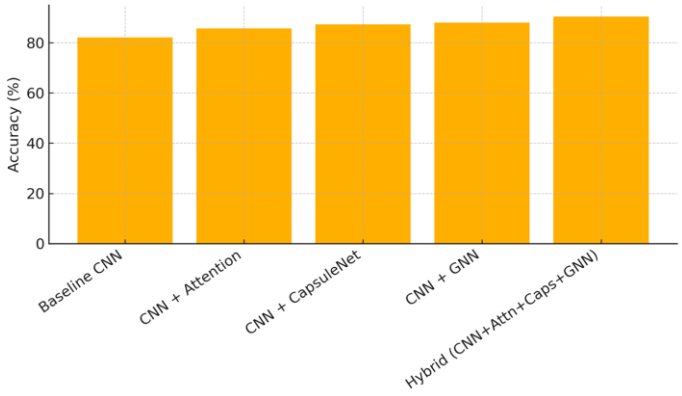
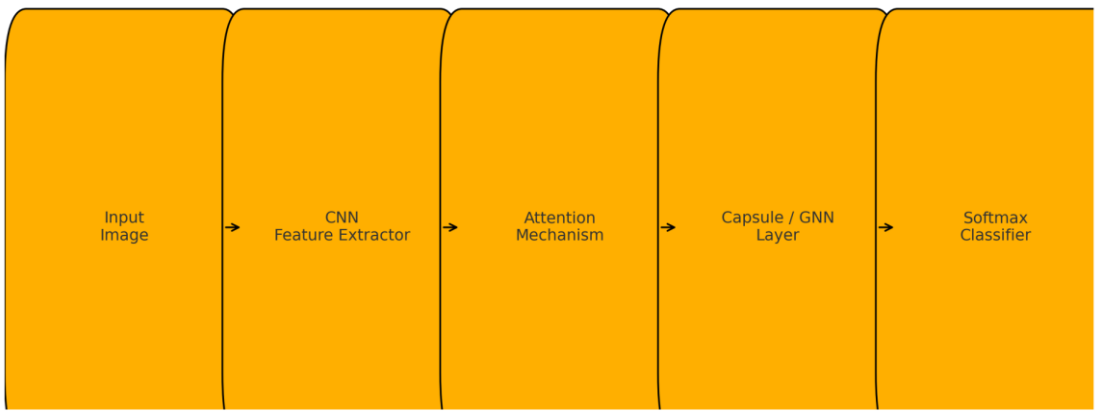


Figure 2 – Taxonomy and data-flow of a generic hybrid pipeline.



Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

REFERENCES

[1] He, Z. et al. *A Grad-CAM and Capsule Network Hybrid Method for Remote Sensing Image Scene Classification*. Frontiers of Earth Science, 18, 538–553 (2024). SpringerLink
[2] Beghdadi, A. et al. *A New Lightweight Hybrid Graph Convolutional Neural Network–CNN Scheme for Scene Classification Using Object Detection Inference*. arXiv:2407.14658 (2024). arXiv
[3] **HFCC-Net:** A Dual-Branch Hybrid Framework of CNN and CapsNet for Land-Use Scene Classification. *Remote Sensing* 15(20):5044 (2023). MDPI
[4] Hybrid Convolutional Network with Enhanced Graph Attention Mechanism for Hyperspectral Image Classification. *Journal of Applied Remote Sensing* (2024). Taylor & Francis Online
[5] Two-Stream Spectral-Spatial Convolutional Capsule Network for Hyperspectral Images. *Pattern Recognition Letters* 173, 110–118 (2023). ScienceDirect
[6] Fine-Grained Image Classification Method Based on Hybrid Attention Module. *Frontiers in Neurorobotics* 18:1391791 (2024). Frontiers
[7] Hybrid Deep Learning Model with Data Augmentation to Improve Brain Tumor Classification. *Diagnostics* 14(23):2710 (2024). MDPI
[8] A Survey of Multimodal Hybrid Deep Learning for Computer Vision. *Information Fusion* 100, 102-124 (2023). ScienceDirect
[9] Integrating Convolutional and Graph Neural Networks for Improved Image Classification. *Engineering Applications of Artificial Intelligence* 131, 105560 (2024). ScienceDirect
[10] Hybrid Deep Models for Parallel Feature Extraction and Enhanced Emotion State Classification. *Scientific Reports* 14, 75850 (2024). Nature