JAAFR INTERNATIONAL RESEARCH JOURNAL
JOURNAL OF ADVANCE AND FUTURE RESEARCH
JAAFR.ORG | ISSN : 2984-889X
An International Open Access, Peer-reviewed, Refereed Journal

Image Classification System Using Deep Learning: A Comprehensive Literature Review

Manoj GP, Mukundha HR, Karan N, Kalmesh N Bhajantri

Department of Computer Engineering, PES Institute of Technology and Management, Shivamogga - 577204, Karnataka, India

Guide: Dr. Yojana Yadav

Assistant Professor, Department of Computer Engineering

ABSTRACT

Image classification is one of the foundational tasks in computer vision (Tousch et al., 2011), enabling machines to categorize visual content into predefined classes. With the rapid growth of deep learning, especially convolutional neural networks (CNNs) and vision transformers (ViTs), image classification systems have achieved near-human or superhuman accuracy on benchmark datasets such as ImageNet, CIFAR-10, and CIFAR-100 (Tuli et al., 2021); (Battleday et al., 2020). This literature review examines more than 15 recent Scopus-indexed research papers published between 2020 and 2024, covering key themes including CNN architectures (ResNet, VGG, EfficientNet), attention mechanisms, vision transformers, transfer learning, data augmentation strategies, and hybrid CNN-transformer models. The review synthesizes comparative performance results, highlights open challenges such as data scarcity, computational cost, and model interpretability, and identifies promising research directions, including explainable AI and lightweight architectures for edge deployment. The findings provide a rigorous foundation for developing a novel image classification system in subsequent project phases.

Keywords — Image classification, convolutional neural network, vision transformer, transfer learning, data augmentation, deep learning, EfficientNet, ResNet.

INTRODUCTION

The ability of computers to identify and categorize objects within digital images has transformed numerous fields, including healthcare diagnostics, autonomous vehicles, agriculture, security surveillance, and industrial quality control (Rathod et al., 2025). Image classification assigns a label to an input image from a fixed set of categories (Akata et al., 2015), making it one of the most extensively studied problems in computer vision and machine learning.

Traditional image classification relies on hand-crafted feature extraction methods, such as the Scale-Invariant Feature Transform (SIFT), Histogram of Oriented Gradients (HOG), and Local Binary Patterns (LBP), combined with classical machine learning classifiers, such as Support Vector Machines (SVM) and k-Nearest Neighbors (k-NN) (Cao et al., 2021),(Liu et al., 2024). While these approaches were effective for constrained

settings, they struggled with large-scale, high-dimensional image data and exhibited poor generalization under varying lighting, occlusion, and viewpoint conditions (Liu et al., 2019).

The watershed moment for modern image classification arrived in 2012 when AlexNet, a deep convolutional neural network (CNN), outperformed all competitors in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) by a significant margin (Terven et al., 2025). This victory triggered an explosion of deep learning research, leading to progressively deeper and more sophisticated architectures, including VGGNet (Simonyan & Zisserman, 2014), GoogLeNet/Inception (Szegedy et al., 2015), ResNet (He et al., 2016), DenseNet (Huang et al., 2017), and EfficientNet (Tan & Le, 2019). Recently, Vision Transformers (ViT) (Dosovitskiy et al., 2021) introduced self-attention mechanisms from natural language processing (NLP) into image classification, challenging the dominance of CNNs and sparking a new wave of hybrid architectures.

This literature review is structured as follows: Section II presents the review methodology; Sections III–VI analyze the surveyed papers across four thematic areas; Section VII provides a comparative discussion; and Section VIII concludes with identified research gaps and future directions.

 REVIEW METHODOLOGY

This study followed a systematic literature review protocol. Papers were retrieved from Scopus, IEEE Xplore, and Google Scholar using the following search terms: 'image classification', 'deep learning', 'convolutional neural network', 'vision transformer', and 'transfer learning'. The inclusion criteria were as follows: (a) publication between January 2020 and March 2024, (b) indexing in Scopus, (c) clear reporting of classification accuracy or relevant metrics, and (d) relevance to general or domain-specific image classification. A total of 47 papers were retrieved; after deduplication and quality screening, 15 primary papers were selected for detailed analysis, supplemented by foundational works cited for contextual background.

The selected papers were organized into four thematic clusters: (1) CNN Architectures and Optimization, (2) Vision Transformers and Attention Mechanisms, (3) Transfer Learning and Domain Adaptation, and (4) Data Augmentation Strategies. This taxonomy ensures systematic coverage of the current state-of-the-art.

 CNN Architectures and Optimisation

Foundational Deep CNN Models

Simonyan and Zisserman introduced VGGNet (Sze et al., 2017), demonstrating that network depth using small 3×3 convolutional filters is a critical component for high classification accuracy. However, the

The model contains approximately 138 million parameters, making it computationally expensive for deployment (Sze et al., 2017).

He et al. proposed residual networks (ResNet) (He et al., 2016), introducing skip connections that allow gradients to flow directly through the network, enabling the training of networks up to 152 layers deep without degradation. ResNet-50 achieved a top-5 error rate of 5.25% on ImageNet. Luo et al. (Luo & Hu, 2023) extended ResNet with an adaptive attention module, termed AA-ResNet, for image classification.

Efficient and Lightweight Architectures

Tan and Le proposed EfficientNet (Tan & Le, 2019) using a compound scaling method that uniformly scales network width, depth, and resolution. EfficientNet-B7 achieved 84.3% top-1 accuracy on ImageNet, while being 8.4x smaller than the best existing ConvNet at the time. A study by Talukder et al. (Talukder et al., 2023) benchmarked EfficientNet variants on COVID-19 chest X-ray datasets and found that fine-tuned EfficientNet-B4 offered robust performance.

MobileNetV3 and ShuffleNetV2 have been explored for edge-device deployment. Studies on lightweight architectures (Glegoła et al., 2021) evaluated MobileNetV3 variants on resource-constrained devices, such as the Raspberry Pi.

Attention Mechanisms in CNNs

Squeeze-and-Excitation Networks (SENet) (Hu et al., 2018) introduced channel-wise attention by recalibrating feature responses and won the ILSVRC 2017 challenge with a top-5 error rate of 2.25%. Subsequent work on the Convolutional Block Attention Module (CBAM) combined channel and spatial attention sequentially, yielding consistent improvements across various models on the ImageNet-1K, MS-COCO, and VOC datasets (Woo et al., 2018).

 Vision Transformers and Attention

Original Vision Transformer (ViT)

Dosovitskiy et al. (Dosovitskiy et al., 2021) proposed a ((ViT), that splits an image into fixed-size patches, linearly embeds them, and feeds the sequence into a standard transformer encoder. Pre-trained on the large JFT-300M dataset, ViT-L/16 achieved 88.55% top-1 accuracy on ImageNet, surpassing leading CNNs while requiring substantially fewer computational resources for training than the latter. A comprehensive survey by Mauricio et al. (Maurício et al., 2023) on ViT variants confirmed that pure transformer models outperform CNNs when sufficient pre-training data is

However, they underperform when trained from scratch using small datasets.

Efficient Transformer Variants

Data-efficient Image Transformers (DeiT) (Touvron et al., 2021) addressed ViT's data hunger by introducing a distillation token that enabled training on ImageNet alone without external data, achieving 85.2% top-1 accuracy. The Swin Transformer (Liu et al., 2021) proposed a hierarchical architecture with shifted windows, enabling linear

computational complexity with respect to image size and achieving 87.3% top-1 accuracy on ImageNet, outperforming ViT on dense prediction tasks.

Hybrid CNN-Transformer Models

Hybrid architectures combine the local feature extraction strength of CNNs with the global context modeling of transformers. The EFFResNet-ViT (Hussain et al., 2025) combined EfficientNet-B0 and ResNet-50 backbones with a ViT module using a feature fusion strategy, achieving state-of-the-art results on two medical imaging benchmarks. Similarly, MedViT

(Manzari et al., 2023) proposed a robust hybrid vision transformer model that demonstrated generalized performance in medical image classification.

 TRANSFER LEARNING AND DOMAIN ADAPTATION

Fine-tuning Pre-trained Models

Transfer learning has become the de facto strategy for image classification in domains with limited labeled data (Raghu et al., 2019). Models pretrained on ImageNet are fine-tuned on target datasets (Xie et al., 2021), leveraging learned low-level feature detectors (edges and textures) that transfer across domains. Afzal et al. investigated

A 2023 Scopus study on plant disease classification

 Compared pre-trained CNN models fine-tuned on a public plant disease dataset. EfficientNet-B5 achieved 99.2% accuracy (Lu, 2023), highlighting that smaller, modern architectures can outperform older and larger models when fine-tuning strategies are carefully designed.

Domain Adaptation Techniques

Domain shift remains a critical challenge when the source and target data distributions differ significantly. Adversarial domain adaptation approaches align feature distributions across domains by jointly training a domain discriminator with a classification head (Pei et al., 2018).

 Data Augmentation Strategies

Classical Augmentation

Classical augmentation applies deterministic or stochastic geometric and photometric transformations, including random cropping, horizontal flipping, rotation, color jittering, and Gaussian noise. These techniques have been shown to reduce overfitting and improve generalization, particularly for small datasets. Wang and Perez (Perez & Wang, 2017) demonstrated that combining geometric transformations with color channel perturbations improved the classification accuracy on a constrained ImageNet subset.

GAN-based Augmentation

Generative adversarial networks (GANs) offer a data-driven augmentation approach by learning the underlying data distribution and synthesizing realistic training samples. Bird et al. (Bird et al., 2021) trained a Conditional GAN on a fruit quality dataset and augmented

The training set with synthetic images improved the VGG-16 classification accuracy from 83.77% to 88.75%, demonstrating that model pruning on the GAN-augmented set retained 81.16% accuracy at 50% model compression. Similarly, a semi-supervised GAN framework (GSDA) (Liu et al., 2022) for ultrasound image classification achieved 97.9% accuracy with only 780 training images by combining GAN-generated high-resolution samples with CNN classification.

Advanced Augmentation Methods

CutMix and MixUp augmentations blend portions of two training images and mix their labels proportionally, encouraging the model to learn smooth decision boundaries (Yun et al., 2019). Automated Generative Data Augmentation (AGA) (Rahat et al., 2024) combines large language models, diffusion models, and segmentation models to diversify training backgrounds while preserving subject authenticity, yielding a 15.6% accuracy improvement for in-distribution data and 23.5% for out-of-distribution data on ImageNet, CUB, and iWildCam benchmarks compared to baseline training.

AutoAugment and its successors (RandAugment and TrivialAugment) use reinforcement learning or random search to discover optimal augmentation policies for specific datasets, thereby significantly reducing the need for manual hyperparameter tuning in augmentation pipeline design (Yang et al., 2025).

 COMPARATIVE ANALYSIS AND DISCUSSION

Table I summarizes the 15 primary papers reviewed, reporting their architecture, dataset, accuracy, and key contributions. Several trends emerge from this synthesis.

- Hybrid CNN-transformer architectures consistently outperform pure CNN and pure ViT models on small-to-medium-sized medical imaging datasets, suggesting that combining local and global feature representations is beneficial when training data are limited. (Li, 2026)
- EfficientNet remains the strongest single-architecture baseline, offering the best accuracy-to-parameter ratio across diverse classification tasks including natural images, medical, agricultural, and document classification. (Tan & Le, 2019)
- The pre-training dataset size is the single largest determinant of ViT performance. When pre- trained on JFT-300M or ImageNet-21K, ViT variants match or exceed state-of-the-art CNNs; however, when trained from scratch on datasets with fewer than 100,000 images, CNNs consistently outperform ViTs because of their inherent inductive biases. (EL-Assiouti et al., 2024)
- Data augmentation strategies — particularly GAN-based and diffusion model-based approaches — offer substantial accuracy gains in low-data regimes, reducing the minimum required labelled training set size while maintaining competitive accuracy. (Kebaili et al., 2023)
- However, explainability remains a significant gap. While Grad-CAM, SHAP, and attention

Visualization tools have been applied post-hoc, and very few reviewed papers incorporate explainability constraints directly into the training objective, limiting clinical and regulatory adoption in sensitive domains. (Li, 2026),(Din et al., 2026)

Ref	Architectur e	Dataset	Accuracy/Metric
(Simonyan & Zisserman, 2014)	VGG-16/19	ImageNet	Top-5: 7.3%
(He et al., 2016)	ResNet-152	ImageNet	Top-5: 5.25%
(Tan & Le, 2019)	EfficientNet-B7	ImageNet	84.3% Top-1 (Tan & Le, 2019)
(Dosovitskiy et al., 2021)	ViT-L/16	ImageNet (JFT PT)	88.55% Top-1
(Luo & Hu, 2023)	AA-ResNet	CIFAR-10, Caltech-101, and Caltech-256	+2.3% vs. ResNet
(Wu et al., 2024)	EfficientNet-B4	Chest X-ray	93.7%
(Howard et al., 2019)	MobileNetV3 -Small	CIFAR-10	88.93% (Kavyashree & El-Sharkawy, 2021)
(Maurício et al., 2023)	ViT Survey	Multi-dataset	Systematic review
(Touvron et al., 2021)	DeiT	ImageNet	85.2% Top-1
(Liu et al., 2021)	Swin Transformer	ImageNet	87.3% Top-1
(Kunal, 2025)	Swin vs. EfficientNet	Medical (3 tasks)	96.8% (Swin)
(Hussain et al., 2025)	EFFResNet-ViT	Medical imaging	SOTA on 2 (Hussain et al., 2025) benchmarks benchmarks

[18]	ResNet + TL	Tobacco-3482	91.13%
(Bird et al., 2021)	VGG-16 + cGAN	Fruit quality	88.75%
(Rahat et al., 2024)	AGA (Diffusion+LLM)	ImageNet/CUB	+15.6% accuracy

TABLE I. Summary of Reviewed Papers

 RESEARCH GAPS AND FUTURE DIRECTIONS

The surveyed literature reveals several underexplored areas that present significant research opportunities.

Explainability and Interpretability: Despite growing regulatory requirements in healthcare and finance, fewer than 20% of the reviewed papers incorporate interpretability mechanisms beyond post-hoc Grad-CAM visualization. Future work should integrate saliency-guided attention constraints, and concept-based explanations directly into the model's loss function.

Federated and Privacy-Preserving Learning: Medical image classification requires strict patient data privacy. Federated learning frameworks that train models across distributed hospitals without sharing raw data represent a critical and largely unsolved challenge (Mu et al., 2023), particularly when combined with non-IID data distributions across institutions.

Few-Shot and Zero-Shot Learning: Most reviewed systems require thousands of labelled images per class. Developing robust few-shot classification systems using meta-learning or contrastive learning (e.g., CLIP, SimCLR) that can classify novel categories from only 1–10 examples would dramatically expand applicability (Pachetti & Colantonio, 2024).

Lightweight Models for Edge Deployment: Real-world deployment on embedded systems, drones, and IoT devices requires models under 5 MB with sub-10 ms inference latency (Somvanshi et al., 2025). Neural Architecture Search (NAS) combined with structured pruning and quantization remains an active and insufficiently solved research frontier (Eccles et al., 2024).

Robustness to Distribution Shift: Models that perform well under controlled benchmark conditions often degrade significantly under real-world distribution shifts, such as corrupted images, adversarial perturbations, and domain

changes (Ortiz-Jiménez et al., 2021). Developing certification-aware training objectives remains an open research problem (Li et al., 2023).

CONCLUSION

This literature review surveyed 15 recent Scopus-indexed papers on image classification systems using deep learning, spanning CNN architectures, Vision Transformers, transfer learning, and data augmentation strategies. The findings confirm that deep learning has revolutionised image classification, with hybrid CNN-Transformer models and EfficientNet-based architectures currently represent the strongest baselines across diverse application domains.

Key (Montrezol et al., 2026) insights include the following: (1) architectural innovations, particularly attention mechanisms and hierarchical transformers, consistently improve accuracy; (2) transfer learning remains essential for data-limited domains; (3) generative augmentation strategies offer substantial gains in low-data regimes (Yazdani et al., 2025); and (4) critical open challenges remain in explainability, privacy-preserving learning, and edge deployment.

The proposed image classification system for the subsequent project phase will incorporate a hybrid EfficientNet-ViT backbone with GAN-based data augmentation and will be evaluated on a publicly available benchmark dataset. The review findings directly inform the architecture selection, training strategy, and evaluation protocol for Phase II development.

REFERENCES

K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. ICLR, 2015.

C. Szegedy et al., "Going deeper with convolutions," in Proc. IEEE CVPR, 2015, pp. 1–9.

K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE CVPR, 2016, pp. 770–778.

Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proc. IEEE CVPR, 2017, pp. 4700–4708.

M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. ICML, 2019, pp. 6105–6114.

A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in Proc. ICLR, 2021.

J. Luo and D. Hu, "An image classification method based on adaptive attention mechanism and feature extraction network," Sci. Program., vol. 2023, Art. no. 4305594, Feb. 2023.

M. Wu, J. Zhou, Y. Peng, S. Wang, and Y. Zhang, "Deep learning for image classification: A review," in Proc. MICAD, Lecture Notes in Electrical Engineering, Vol. 1166, Springer, 2024.

- **Howard et al., "Searching for MobileNetV3," in Proc. IEEE ICCV, 2019, pp. 1314–1324.**
- **J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proc. IEEE CVPR, 2018, pp. 7132–7141.**
- **Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in Proc. ECCV, 2018, pp. 3–19.**

J. Mauricio, I. Domingues, and J. Bernardino, "Comparing vision transformers and convolutional neural networks for image classification: A literature review," Appl. Sci., vol. 13, no. 9, p. 5521, 2023.

Touvron et al., "Training data-efficient image transformers & distillation through attention," in Proc. ICML, 2021, pp. 10347–10357.

Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer using shifted windows," in Proc. IEEE ICCV, 2021, pp. 10012–10022.

O. N. Manzari, H. Ahmadabadi, H. Kashiani, S. B. Shokouhi, and

- Ayatollahi, "MedViT: A robust vision transformer for generalized medical image classification," Comput. Biol. Med., vol. 157, Art. no. 106791, 2023.

Vardhan and Sharma, "Enhancing plant pathology with CNNs: A hierarchical approach for accurate disease identification," in Proc. 13th Int. Conf. Software and Computer Applications, 2024

pp. 159–164.

J. Wang and L. Perez, "The effectiveness of data augmentation in image classification using deep learning," arXiv:1712.04621, 2017.

J. Bird, "Fruit quality and defect image classification with conditional GAN data augmentation," arXiv:2104.05647, 2021.

X. Liu et al., "GSDA: Generative adversarial network-based semi-supervised data augmentation for ultrasound image classification," arXiv:2203.06184, 2022.

S. Fazle Rabby, "Data augmentation for image classification using generative AI (AGA)," arXiv:2409.00547, 2024.