

Class Imbalance in Deep Learning-Based Image Classification: A Systematic Review

Mochamad Arief Saputra^{1*}

* Teknik Informatika, STMIK IKMI Cirebon
ariefsaputra220@gmail.com¹

Article Info

Article history:

Received 2026-08-10

Revised 2026-10-09

Accepted 2026-10-17

Keyword:

*Class Imbalance,
Deep Learning,
Image Classification,
Imbalanced Learning,
Systematic Literature Review.*

ABSTRACT

Class imbalance is a major challenge in deep learning-based image classification because unequal class distributions can bias model learning toward majority classes and reduce minority-class recognition. This study conducted a Systematic Literature Review (SLR) to identify, classify, and synthesize recent approaches for handling class imbalance in deep learning-based image classification. The review followed the PRISMA 2020 framework and used Scopus as the primary bibliographic database, with publications from 2022 to 2026. A total of 1,630 records were initially retrieved, and 34 articles underwent full-text assessment. After applying the predefined eligibility criteria, 31 studies were included in the final thematic synthesis. The studies were organized into four methodological themes: Traditional Data-Level Approaches, Algorithm-Level and Loss Function-Based Approaches, Generative and Hybrid Approaches, and Emerging Architecture-Integrated Approaches. The synthesis showed that traditional methods such as SMOTE, ADASYN, and augmentation remained useful, while algorithm-level, generative, and architecture-integrated approaches provided alternative mechanisms for addressing minority-class learning. No single approach consistently demonstrated superior performance across different datasets and experimental settings. Effectiveness was influenced by dataset characteristics, class distribution, architecture, evaluation procedures, and computational requirements. The review identifies research gaps in cross-dataset validation, standardized benchmarking, minority-sensitive evaluation, computational-efficiency reporting, and validation of synthetic-data quality and its effect on classification performance.



This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.

I. INTRODUCTION

The development of Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL) has substantially expanded the application of automated image classification in medical imaging, agriculture, industrial inspection, remote sensing, and other visual recognition domains. Deep learning architectures such as Convolutional Neural Networks (CNNs), residual networks, Vision Transformers (ViTs), and transfer learning have demonstrated strong capabilities in learning discriminative visual representations for image classification tasks [1], [2].

Despite these developments, class imbalance remains an important challenge in deep learning-based image classification. Class imbalance occurs when the number of observations differs substantially among classes, causing the learning process to be dominated by majority-class samples. This condition may result in biased classification performance and inadequate recognition of minority classes, particularly

when conventional evaluation measures are used without considering the class distribution [3], [4].

Several approaches have been developed to address this problem. Data-level approaches modify the distribution or composition of training data through oversampling, undersampling, and augmentation. Algorithm-level approaches modify the learning objective through class weighting, cost-sensitive learning, or specialized loss functions. Generative approaches produce additional minority-class samples, while hybrid approaches combine multiple imbalance-handling mechanisms. These strategies have been investigated across different datasets, application domains, and deep learning architectures [5], [3].

However, the increasing number of approaches has made the available evidence more difficult to interpret across studies. Individual studies frequently evaluate a particular imbalance-handling technique using a specific dataset, class distribution, application domain, and deep learning architecture. Consequently, reported improvements cannot

always be compared directly because differences in datasets, imbalance ratios, architectures, and evaluation metrics may influence the observed results. In imbalanced classification, accuracy alone may also provide an incomplete representation of model performance, making minority-sensitive measures such as recall, F1-score, balanced accuracy, G-Mean, and AUC relevant to the interpretation of results [6], [7].

Another issue concerns the relationship between the imbalance-handling mechanism and the deep learning architecture. CNNs, ViTs, transfer learning, ensemble models, and other architectures should not automatically be considered imbalance-handling techniques. Their relevance depends on how they interact with a strategy designed to improve minority-class representation, modify the learning objective, or improve feature representation. This distinction is important because increasing architectural complexity may improve representation capacity while also increasing computational requirements. Therefore, this review distinguishes the mechanism used to address class imbalance from the deep learning architecture used to implement or support that mechanism.

Although research on imbalanced learning continues to expand, the available evidence remains fragmented across datasets, application domains, imbalance-handling mechanisms, and deep learning architectures. Existing studies provide important individual contributions, but their findings are frequently reported under different experimental conditions. This fragmentation limits direct cross-study comparison and makes it difficult to understand how dataset characteristics, class distribution, architecture, evaluation metrics, and computational requirements influence the reported effectiveness of different imbalance-handling strategies.

Therefore, a systematic synthesis is needed to organize existing approaches according to their dominant imbalance-handling mechanism and to examine the factors associated with their reported effectiveness. In addition to classification performance, the synthesis considers minority-class recognition, generalization, computational requirements, and the distinction between the generation of synthetic data and the resulting classification improvement.

Accordingly, this study conducts a Systematic Literature Review (SLR) to identify and synthesize recent approaches for handling class imbalance in deep learning-based image classification. The study aims to identify the principal techniques, organize them into methodological themes, compare their reported effectiveness under different experimental contexts, examine factors influencing their performance, and identify remaining research gaps and future directions.

The review addresses four research questions: RQ1. What techniques are used to handle class imbalance in deep learning-based image classification? RQ2. How effective are data-level, algorithm-level, generative, and hybrid approaches in improving classification performance? RQ3. What factors influence the effectiveness of class-imbalance handling methods? RQ4. What research gaps remain, and

what future directions can be identified for addressing class imbalance in deep learning-based image classification.

The main contribution of this study is a thematic synthesis that organizes imbalance-handling approaches according to the dominant stage at which the imbalance problem is addressed: training-data distribution, model optimization, minority-data generation, and integration with deep learning architectures. The review further examines factors associated with the reported effectiveness of these approaches and identifies research gaps across the four methodological themes. This synthesis provides a structured interpretation of the literature while considering the trade-offs among classification performance, minority-class representation, model complexity, and computational requirements.

II. METHODS

A. Research Design and Search Strategy

This study employed a Systematic Literature Review (SLR) to identify, evaluate, and synthesize research concerning class imbalance handling in deep learning-based image classification. The review process was structured according to the PRISMA 2020 framework to support transparent identification, screening, eligibility assessment, and inclusion of relevant studies [8].

Scopus was used as the primary bibliographic database for the literature search. The search was conducted on 28 July 2026 and was limited to publications from 2022 to 2026. The search strategy combined terms representing class imbalance, image classification, deep learning, and imbalance-handling techniques. The Boolean search string used in Scopus was: ("class imbalance" OR "imbalanced data" OR "imbalanced dataset" OR "long-tailed learning" OR "minority class") AND ("image classification" OR "image dataset" OR "computer vision" OR "medical image" OR "image recognition") AND ("deep learning" OR "machine learning" OR CNN OR "vision transformer" OR ViT OR "neural network") AND (SMOTE OR oversampling OR undersampling OR augmentation OR "cost-sensitive learning" OR "focal loss" OR "ensemble learning")

The initial search retrieved 1,630 records. The search was restricted to English-language journal articles published during the defined period according to the eligibility criteria. Because Scopus was used as the primary database, studies indexed exclusively in other bibliographic databases may not have been retrieved. This limitation introduces a potential database-selection and publication bias and should be considered when interpreting the findings.

B. Inclusion and Exclusion Criteria

Studies were included when they were peer-reviewed journal articles published in English between 2022 and 2026, addressed image-based classification using deep learning or deep neural network methods, explicitly investigated class imbalance or minority-class learning, evaluated an imbalance-handling strategy, and provided sufficient

methodological and experimental information for thematic synthesis.

Studies were excluded when they were review articles, systematic reviews, book chapters, conference proceedings, editorials, or other non-journal publications; were published in languages other than English; did not use image data; focused primarily on video, text, multimodal text-oriented tasks, physiological signals, or other non-image inputs; did not explicitly address class imbalance; focused only on architecture or model development without evaluating an imbalance-handling strategy; addressed a classification application without investigating the imbalance problem; or did not provide accessible full text or sufficient methodological information for analysis.

C. Screening, Data Extraction, and Synthesis

The screening process was conducted sequentially from identification to final inclusion. Retrieved records were screened using the predefined eligibility criteria, followed by title and abstract screening and full-text assessment. A total of 34 articles were assessed at the full-text stage. Three articles were excluded from the final synthesis because they did not satisfy the operational scope of image-based deep learning classification with explicit class-imbalance handling. The excluded studies were [9], which focused primarily on text-oriented visual question answering [10], which focused on spatio-temporal video classification [11], which focused on ECG signal classification rather than image data.

The remaining 31 studies were included in the final review and thematic synthesis. The previous distinction between 31 included studies and 28 studies used for thematic synthesis was therefore removed. All 31 eligible studies were included in the final thematic analysis.

For each included study, information was extracted concerning the author and publication year, dataset and application domain, class distribution or imbalance ratio when reported, deep learning architecture, imbalance-handling technique, baseline or comparison method, evaluation metrics, experimental results, and computational requirements when available. For generative approaches, information concerning synthetic-data generation and its relationship to classification performance was also extracted.

The 31 included studies were organized into four methodological themes: Traditional Data-Level Approaches; Algorithm-Level and Loss Function-Based Approaches; Generative and Hybrid Approaches; and Emerging Architecture-Integrated Approaches. The themes were defined according to the dominant point of intervention used to address class imbalance. Data-level approaches modify training-data distribution, algorithm-level approaches modify the learning objective or optimization process, generative and hybrid approaches increase or transform minority representations, and architecture-integrated approaches incorporate imbalance handling into or alongside emerging deep learning architectures. Studies involving multiple techniques were classified according to their dominant methodological contribution.

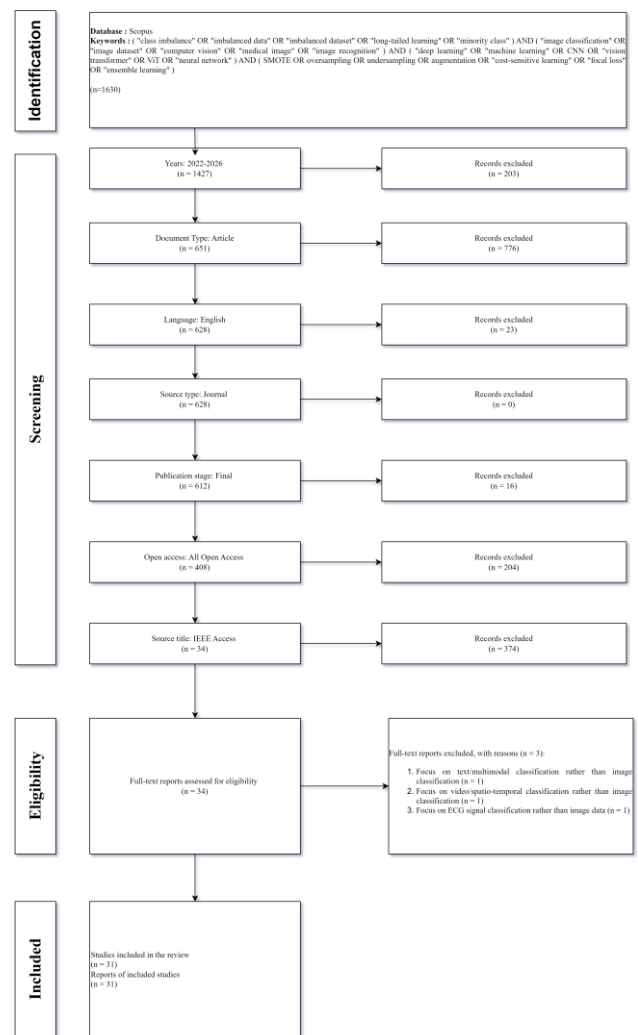


Figure 1. PRISMA 2020 Flow Diagram of the Literature Selection Process

The synthesis considered reported evaluation measures including accuracy, precision, recall, F1-score, AUC, balanced accuracy, G-Mean, and minority-class performance when available. These measures were interpreted within the context of each study because the included studies differed in datasets, class distributions, architectures, and experimental settings. Particular attention was given to minority-sensitive metrics because overall accuracy may not adequately represent performance under class-imbalanced conditions [8].

For generative approaches, the synthesis distinguished between the ability to generate additional or realistic minority samples and the subsequent improvement in classification performance. Consequently, synthetic-data quality was not interpreted as evidence of improved classification performance unless corresponding experimental results were reported.

Computational requirements and model complexity were also considered when these were reported by the original studies. These factors were used to identify potential trade-offs between classification performance and computational cost.

D. Quality Assessment

The methodological quality of the included studies was assessed using a predefined qualitative assessment checklist. The checklist considered eight aspects: (1) relevance of the study to the review objectives, (2) clarity of the research methodology, (3) adequacy of dataset and class-distribution description, (4) clarity of the deep learning architecture and imbalance-handling method, (5) appropriateness of the experimental or validation strategy, (6) completeness and appropriateness of the evaluation metrics, (7) completeness of experimental result reporting, and (8) reproducibility and potential sources of methodological bias. The criteria were applied consistently to the included studies based on the methodological and reporting information available in each article.

The assessment criteria were informed by established quality-assessment perspectives relevant to artificial intelligence and medical or computer vision research, including [12], [13]. These references were used to support the assessment dimensions concerning methodological transparency, reporting quality, validation, reproducibility, and potential sources of bias.

The quality assessment was not used as a performance ranking of the included studies. In particular, a study was not considered to have lower methodological quality merely because it reported lower classification performance. Instead, the assessment was used to identify differences in methodological strength and reporting completeness and to contextualize the interpretation of findings during the synthesis. Studies with limitations in dataset description, validation procedures, evaluation metrics, or experimental reporting were therefore considered more cautiously when comparing their findings with those of other studies.

No numerical minimum-score threshold was applied because the quality assessment was designed as a qualitative appraisal rather than a quantitative scoring instrument. Consequently, studies were not excluded solely on the basis of a quality score. Eligibility was determined using the predefined inclusion and exclusion criteria in Section B, while the quality assessment was subsequently used to evaluate the strength, limitations, and comparability of the evidence included in the thematic synthesis.

The results of the quality assessment were considered when interpreting evidence across the four methodological themes, particularly when comparing reported improvements under different datasets, class-imbalance conditions, deep learning architectures, evaluation metrics, and experimental settings. This approach was intended to prevent reported classification performance from being interpreted independently of the methodological quality and completeness of the underlying evidence.

III. RESULTS AND DISCUSSION

The final thematic synthesis was organized into four methodological themes: Traditional Data-Level Approaches, Algorithm-Level and Loss Function-Based Approaches, Generative and Hybrid Approaches, and Emerging

Architecture-Integrated Approaches. The themes were established according to the primary intervention point through which each study addressed class imbalance. Data-level approaches primarily modified the distribution or diversity of training samples; algorithm-level approaches modified the learning objective or contribution of classes during optimization; generative and hybrid approaches generated or combined additional minority representations; and architecture-integrated approaches combined imbalance-related strategies with contemporary deep learning architectures. When a study employed multiple techniques, it was assigned to the theme representing its dominant methodological contribution to avoid double counting across themes.

The synthesis of the final corpus was used to answer RQ1 concerning the techniques employed to address class imbalance, RQ2 concerning their reported effectiveness, RQ3 concerning the factors influencing effectiveness, and RQ4 concerning remaining research gaps and future directions. Because the reviewed studies differed substantially in datasets, imbalance conditions, architectures, evaluation protocols, and application domains, reported performance values were interpreted within their respective experimental contexts rather than used to establish a direct ranking among methods.

A. Traditional Data-Level Approaches

Traditional data-level approaches address class imbalance by modifying the distribution, quantity, or diversity of training samples. The reviewed studies included SMOTE-based oversampling, ADASYN, and conventional or task-specific image augmentation. These approaches primarily increase the representation of minority classes or introduce additional variation before model optimization.

TABLE I
TRADITIONAL DATA-LEVEL APPROACHES

Author(s) & Year	Proposed Method	Dataset / Application	Main Finding
Fareed et al. (2022)	SMOTE-Tomek + ADD-Net	Kaggle Alzheimer's Brain MRI	98.63% accuracy and 99.76% AUC; reduced class overlap through oversampling and Tomek-link cleaning.
Alshamrani & Alshomrani (2025)	Dual-SMOTE Framework + ResNet50	Mammogram Breast Cancer	Demonstrated that applying SMOTE only to training data provides a more realistic evaluation setting.
Farhan et al. (2025)	OCMRI Data Augmentation	Brain MRI Tumor Datasets	Accuracy increased from 85.19–94.77% to 92.70–98.00%
Mohammed et al. (2023)	ADASYN + Fully	Collective Motion Database	ADASYN-FCNN achieved 93.97% accuracy and was

	Connected DNN		reported to outperform standard SMOTE.
Akhtar et al. (2024)	Data Augmentation + VGG-19 Transfer Learning	Urine Sediment Images	Data augmentation was used to increase minority samples and address class imbalance; the proposed approach achieved 98% accuracy with 61 ms inference time.

[14] employed SMOTE combined with Tomek-link cleaning and ADD-Net for Alzheimer's brain MRI classification. The study reported 98.63% accuracy and 99.76% AUC. The combination of oversampling and Tomek-link cleaning was intended to increase minority representation while reducing overlapping samples between classes. The result indicates that the effectiveness of resampling may depend not only on the number of generated samples but also on the treatment of potentially ambiguous observations.

[15] investigated a Dual-SMOTE framework with ResNet50 for mammogram breast-cancer classification. Their experimental comparison distinguished between applying SMOTE to the complete dataset and applying it only to the training data. This distinction highlights an important methodological issue in imbalanced learning: synthetic samples should be generated within the training data so that evaluation data remain independent of the balancing procedure. Consequently, improvements obtained from balancing the complete dataset should not be interpreted equivalently to improvements measured on an untouched evaluation set.

Image augmentation represented another recurring data-level strategy. [16] proposed Oriented Combination MRI (OCMRI) augmentation for brain MRI tumor datasets and reported accuracy improvements across the evaluated datasets. The method increased training-image variation by combining existing image information rather than introducing a separate generative model. [17] employed a data-centric and model-centric strategy incorporating data augmentation and VGG-19-based transfer learning for multi-class automated urine-sediment classification. The model reported 98% accuracy with an inference time of approximately 61 ms per image. This study is relevant to data-level synthesis because it demonstrates how augmentation can be incorporated into a data-centric pipeline while also considering model performance and inference requirements.

[18] employed ADASYN with a fully connected deep neural network and reported 93.97% accuracy. The study reported better performance than standard SMOTE within its experimental setting. This comparison should be interpreted as within-study evidence rather than evidence that ADASYN is generally superior to SMOTE because the studies across the reviewed literature use different datasets, class distributions, model configurations, and evaluation procedures.

The reviewed data-level methods also differ according to the representation being modified. SMOTE and ADASYN generally construct additional minority representations through feature-space resampling, whereas image augmentation operates directly on image representations. Therefore, the two approaches cannot be compared solely according to the number of additional samples produced. Feature-space interpolation and image-space transformation introduce different forms of variation and may respond differently to the complexity and diversity of minority classes.

Overall, the evidence indicates that traditional data-level approaches remain practical because they can generally be incorporated without fundamentally changing the classifier. Their effectiveness, however, depends on factors including imbalance severity, minority-class diversity, quality of the original samples, and the separation between training and evaluation data. Increasing the number of training samples does not necessarily improve minority-class recognition when the generated or transformed samples provide limited new information.

The evaluation metrics also influence the interpretation of these findings. Overall accuracy can remain high when the majority class dominates the evaluation set, even when minority recognition is inadequate. Therefore, when available, macro-F1, balanced accuracy, minority-class recall, G-Mean, and AUC provide complementary evidence for evaluating imbalance-handling procedures [6], [7].

B. Algorithm-Level and Loss Function-Based Approaches

Algorithm-level approaches address class imbalance by modifying how classes or individual samples contribute to model optimization rather than primarily changing the number of training samples. The reviewed studies included class weighting, focal-loss mechanisms, balanced-loss strategies, and optimization procedures designed to reduce the dominance of majority classes during training.

TABLE II
ALGORITHM-LEVEL AND LOSS FUNCTION-BASED APPROACHES

Author(s) & Year	Proposed Method	Dataset / Application	Main Finding
Hoang et al. (2023)	Categorical Weighting Domination (CWD)	ISIC2018 Skin Cancer	Recall increased by 5.16% without oversampling.
Nazih et al. (2023)	ViT + Class Weights + Focal Loss	FGADR Diabetic Retinopathy	The combination of focal loss and class weighting produced an F1-score of 0.825 on a highly imbalanced dataset.
Karahan et al. (2026)	YOLOv11m + Class-Balanced Focal Loss + Canny Edge Maps	Plant Phenology Dataset	Reported 93.97% mAP@0.5 and emphasized minority-class samples through

			class-balanced focal loss.
Ali et al. (2024)	CNN + Tangent Optimization Algorithm (TOA)	B-Mode Ultrasound Carotid Plaque Images	TOA optimized CNN hyperparameters for imbalanced data without preprocessing or augmentation; achieved 98.82% accuracy, 99.41% sensitivity, and 95.74% specificity.

[19] proposed Categorical Weighting Domination (CWD), an adaptive weighting mechanism for imbalanced skin-cancer classification. The study reported an increase in recall without relying on oversampling. This result illustrates how the learning objective can be modified so that minority observations receive greater influence during optimization while the original training distribution is retained.

[20] combined a Vision Transformer with class weighting, focal loss, label smoothing, and AdamW optimization for diabetic-retinopathy classification. The reported results included 82.6% balanced accuracy, 96.4% AUC, and an F1-score of 82.5%. The use of balanced accuracy and AUC provides information beyond overall accuracy and is particularly relevant when class distributions are unequal.

[21] employed class-balanced focal loss with YOLOv11m and Canny edge maps for plant-phenology analysis. The study reported 93.97% mAP@0.5. The study illustrates that loss-level balancing can also be integrated into contemporary vision architectures in which minority representation influences the optimization of object-level predictions.

[22] proposed CAROTIDNet combined with the Tangent Optimization Algorithm (TOA) for carotid-plaque ultrasound classification. The study reported 98.82% accuracy, 99.41% sensitivity, and 95.74% specificity. The study illustrates an algorithm-level direction in which optimization is integrated into the classification pipeline under an imbalanced medical-image setting rather than relying primarily on synthetic sample generation.

[23] combined a multimodal Vision Transformer, ensemble learning, improved Grey Wolf Optimization, and SMOTE for battery-usage-profile classification. Because several imbalance-related mechanisms were integrated, the study demonstrates the methodological overlap between data-level and algorithm-level strategies. For the thematic synthesis, its dominant contribution is considered in relation to optimization and ensemble mechanisms rather than treating SMOTE independently from the complete framework.

Across these studies, algorithm-level methods shift the intervention point from the training-data distribution toward the optimization process. Their effectiveness remains dependent on class distribution, architecture, hyperparameter selection, and evaluation protocol. The reviewed evidence therefore does not establish that focal loss, class weighting,

balanced loss, or optimization-based approaches consistently outperform other imbalance-handling strategies across different image-classification settings.

For RQ2, this finding indicates that effectiveness should be interpreted primarily through within-study comparisons against appropriate baselines. A high value of one metric in one dataset does not constitute evidence of general superiority over another method evaluated under a different dataset and experimental protocol.

C. Generative and Hybrid Approaches

Generative and hybrid approaches extend imbalance handling by generating additional minority representations or combining generation with conventional balancing techniques. The reviewed studies included GAN-based methods, conditional generation, VAE-based approaches, diffusion-based generation, and hybrid combinations involving feature-space oversampling.

TABLE III

GENERATIVE AND HYBRID APPROACHES

Author(s) & Year	Proposed Method	Dataset / Application	Main Finding
Qin et al. (2022)	Progressive GAN + Multi-Scale CNN	Chest X-ray Images	Generated realistic synthetic medical images for small and imbalanced datasets.
Su et al. (2024)	Self-Transfer GAN (STGAN)	HAM10000 Skin Lesion	98.23% accuracy and improved minority-class synthesis.
Kumar et al. (2025)	DiSMHA (Diffusion + Feature SMOTE)	Solar Panel Dust Detection	Reported 96.67% accuracy for the MLP and combined pixel-level generation with feature-level balancing.
Sreelekshmi et al. (2024)	Two-Point Crossover Augmentation + Fourth-order Complex Diffusion + HVAE	MIAS & KAU-BCMD	Reported 99.80% accuracy on KAU-BCMD, with AUC of 99.30%, balanced accuracy of 99.80%, and MCC of 99.20%; the approach combined augmentation, multi-scale diffusion representation, and hierarchical VAE feature extraction.
Kovac et al. (2026)	Unconditional StyleGAN2-ADA	IDR1D & FGADR	Generated stage-specific synthetic diabetic-retinopathy images with emphasis on clinical realism.
Alam et al. (2025)	Stable Diffusion-Based	Forest Fire Image Dataset	Reported 99.05% accuracy and

	Augmentation		reduced false-positive rate.
Asokan & Manju (2026)	Mask-Conditioned GAN + Lesion Fidelity-Aware Filtering	Indian DR Dataset	Synthetic-image filtering based on lesion consistency improved ordinal classification agreement relative to random augmentation
Park et al. (2022)	Feature Dictionary-Based Minority Oversampling	Public Fashion Image Datasets	A feature-dictionary-based generative model synthesized minority samples to balance training; experiments reported the highest top-1 performance on several public fashion datasets.
Shafique et al. (2025)	CGAN-Based Data Augmentation + ConvNeXt-Sim/SimAM	Hyperspectral Apple Images	CGAN generated category-specific synthetic images to address class imbalance; the proposed model achieved 99.75% accuracy.

[24] investigated Progressive GAN combined with a multi-scale CNN for small and imbalanced chest X-ray datasets. The study focused on the generation of additional medical-image representations, illustrating that the usefulness of synthetic data depends on whether generated samples preserve characteristics relevant to the downstream classification task.

[25] developed a Self-Transfer GAN (STGAN) for HAM10000 skin-lesion images. The approach addressed minority-class synthesis and reported 98.23% accuracy and an 89.48% F1-score. The study also considered generation-oriented measures such as FID and Inception Score. These measures provide information about synthetic-image quality, whereas classification metrics measure downstream recognition performance; the two outcomes therefore need to be interpreted separately.

[26] proposed DiSMHA, which combined diffusion-based image augmentation with feature-level SMOTE for solar-panel dust detection. The method operated at both image-space and feature-space levels. This provides an example of a hybrid approach in which synthetic image generation and feature-space oversampling introduce minority information through different mechanisms.

[27] proposed a feature dictionary-based generative framework for minority-class oversampling in imbalanced image classification. The method constructs feature dictionaries from pretrained feature extractors to capture shape and texture information and subsequently generates artificial minority features through adversarial training rather than directly synthesizing raw images. The generated minority features are then used to fine-tune the classifier to

improve class balance. Experiments on public fashion datasets with extreme class imbalance showed that the proposed approach achieved strong top-1 classification performance, while ablation studies were used to examine the contribution of the feature dictionaries and the number of generated samples.

[28] employed a Conditional GAN with ConvNeXt and SimAM for hyperspectral apple images. The framework generated class-specific synthetic hyperspectral representations before classification and reported 99.75% accuracy. The study demonstrates the additional requirement of preserving domain-specific information when synthetic data contain spectral characteristics beyond conventional RGB image information.

[29] employed Stable Diffusion-based synthetic image augmentation for forest-fire classification and reported 99.05% accuracy, 99.41% precision, and 98.28% recall. The findings demonstrate the potential of diffusion-based generation to introduce additional visual variation, although classification performance alone does not establish the fidelity or diversity of the generated images.

[30] employed StyleGAN2-ADA for stage-specific diabetic-retinopathy fundus-image generation and evaluated generation quality using FID and KID together with clinical assessment. [31] introduced lesion-fidelity-aware conditional GAN augmentation and incorporated lesion consistency into the filtering of generated samples. These studies demonstrate a shift toward domain-specific control and validation of synthetic data rather than treating the quantity of generated samples as the primary objective.

[32] combined two-point crossover augmentation, fourth-order complex diffusion, and a hierarchical variational autoencoder. The study reported high classification performance on the evaluated breast-cancer datasets and incorporated AUC, balanced accuracy, and MCC into the evaluation. Because the approach integrates augmentation, diffusion, and VAE-based representation learning, it is categorized according to its dominant generative and hybrid contribution.

The synthesis of these studies requires an explicit distinction between synthetic-data generation capability and classification-performance improvement. A generative model may produce visually realistic or statistically plausible samples without necessarily improving minority-class recognition. Conversely, an increase in classification performance does not by itself demonstrate that generated samples are sufficiently realistic, diverse, or valid for the target application. Generation-quality measures and downstream classification metrics should therefore be interpreted as complementary evidence.

Generative approaches also introduce additional computational and methodological requirements. GANs, diffusion models, and VAE-based approaches require additional training or generation stages, while synthetic samples require validation for fidelity, diversity, and domain consistency. These requirements distinguish generative approaches from conventional resampling and augmentation.

Overall, the reviewed evidence supports generative and hybrid methods as an important direction for increasing minority representation, but it does not establish their universal superiority over conventional data-level or algorithm-level approaches. Their usefulness depends on the quality of the original minority data, the ability of the generative model to preserve task-relevant characteristics, the evaluation protocol, and computational requirements.

D. Emerging Architecture-Integrated Approaches

Emerging architecture-integrated approaches combine imbalance-related mechanisms with contemporary deep learning architectures. The reviewed studies included Vision Transformers, transfer learning, hybrid CNNs, ensemble learning, knowledge distillation, federated learning, and lightweight architectures. In this theme, the architecture itself is considered part of the imbalance-handling strategy when it is integrated with a mechanism relevant to learning under imbalanced conditions.

TABLE IV
EMERGING ARCHITECTURE-INTEGRATED APPROACHES

Author(s) & Year	Proposed Method	Dataset / Application	Main Finding
Veda Sri et al. (2026)	Privacy-preserving federated learning	Retinal Disease Diagnosis	98.90% accuracy using privacy-preserving collaborative learning.
Rocha et al. (2026)	Ensemble Transfer Learning + Threshold-Based Resampling	ISIC Skin Tumor Dataset	Improved robustness and F2-score for melanoma identification.
Mashal et al. (2025)	ViT with Transfer Learning	Skin Disease Dataset	Reported 97.53% accuracy and emphasized attention-based representation of lesion features.
Venkatachalam & Shah (2025)	Dual-Branch Hybrid CNN (Xception + ResNet)	Lung and Colon Cancer	Reported 98.96% accuracy through feature fusion between two CNN branches.
Yamaçlı (2025)	Multimodal ViT + Ensemble Soft Voting	Battery Usage Profile	Reported 92.93% accuracy using multimodal representation and ensemble learning.
Chatterjee et al. (2024)	TL Ensemble (InceptionV3 + Xception)	HAM10000 Skin Lesion	98% accuracy after augmentation and transfer learning.
Prayogo et al. (2025)	Hybrid CNN (ResNet, MobileNet, DenseNet)	Brain MRI Tumor	98.75% accuracy, precision, recall, and F1-score at approximately the same level.

Arora et al. (2026)	Cross-Modality Transfer Learning	BreakHis	Achieved an AUC of 0.9708 through transfer of representations across imaging modalities.
Santos et al. (2025)	Fine-Tuned Multilevel CNN	DFUC	Reported 95.91% accuracy and 93.28% Kappa for diabetic-foot-ulcer classification.
Vemuri et al. (2026)	CB-DistillGrad	Road Surface Condition Dataset	Reduced model parameters substantially while retaining 87.11% accuracy.
Ahmed et al. (2022)	MobileNetV2 + CLAHE + Runtime Augmentation	PlantVillage	Reported 99.30% accuracy with a lightweight model of approximately 9.60 MB.
Hassan Ali et al. (2026)	AttFKAN-DEQ + Lightweight Attention	BreakHis & IDC Histopathology	Combined Fourier KAN, deep equilibrium learning, and attention for histopathology classification; achieved 96.60%/94.52% accuracy and 98.44%/97.31% AUC on BreakHis/IDC.
Diaz et al. (2024)	Pre-trained CNNs + Data Augmentation	Sterile Barrier Packaging Seal Images	Evaluated CNN resilience under highly imbalanced defect datasets; some architectures remained capable of classification when defective samples were as low as 5%.

[33] combined a Vision Transformer with transfer learning for skin-disease image classification and reported 97.53% accuracy with an F1-score of 0.98, together with high sensitivity and specificity. The study demonstrates the use of attention-based representation learning in an imbalanced image-classification setting, while also indicating the computational and pretraining requirements associated with transformer-based architectures.

[26] proposed a dual-branch hybrid CNN combining Xception and ResNet for histopathological lung and colon cancer classification. The model reported 98.96% accuracy through feature fusion. The study illustrates how complementary feature representations can be combined, although the use of multiple backbone networks may increase model complexity and computational requirements.

[34] combined InceptionV3 and Xception through transfer-learning-based ensemble learning for HAM10000 skin-lesion classification. The model reported approximately 98% training and testing accuracy after augmentation and transfer learning. The ensemble framework provides complementary feature representations, but the use of multiple backbones can increase parameter count, memory consumption, and training requirements.

[35] employed hybrid CNN transfer learning using ResNet50V2, MobileNetV2, and DenseNet121 with augmentation for brain MRI tumor classification. The reported accuracy, precision, recall, and F1-score were approximately 98.75%. The study demonstrates how imbalance-related augmentation can be integrated with multiple pretrained CNN backbones rather than being implemented independently from the classification architecture.

[36] employed a fine-tuned multilevel CNN for diabetic-foot-ulcer classification and reported 95.91% accuracy and 93.28% Kappa. The study is retained in this theme because its dominant contribution concerns an architecture-integrated classification framework rather than conventional data-level resampling.

[37] investigated cross-modality transfer learning using representations transferred between imaging modalities for breast-cancer classification. The study reported 92.75% accuracy and an AUC of 0.9708. The finding highlights the potential of transfer learning when target-domain data are limited, while also emphasizing the importance of the domain gap between source and target representations.

[38] proposed privacy-preserving federated learning with homomorphic encryption and transfer learning for retinal disease diagnosis and reported 98.90% accuracy. The study represents an emerging direction in which image classification is considered together with privacy and distributed-learning requirements. Such a framework also introduces additional communication and computational requirements.

[39] introduced CB-DistillGrad using class-balanced distillation and gradient-conflict resolution for lightweight road-surface classification. The model achieved 87.11% accuracy with approximately 130K parameters. This study is relevant to deployment-oriented analysis because model size is considered together with classification performance.

[40] combined ensemble transfer learning with threshold-based resampling for skin-tumor classification. The reported findings emphasized robustness and F2-score for melanoma identification. The study illustrates that architecture-level strategies can be combined with decision-threshold or resampling mechanisms rather than relying on a single imbalance intervention.

[41] combined MobileNetV2, CLAHE, and runtime augmentation for PlantVillage classification and reported 99.30% accuracy using a lightweight model of approximately 9.60 MB. The study is relevant to deployment-oriented analysis because model size and preprocessing are considered together with classification performance.

[42] introduced AttFKAN-DEQ, an architecture that combines Fourier Kolmogorov-Arnold Networks (FKANs), Deep Equilibrium Models (DEQs), and a lightweight convolutional block attention module (LCBAM) for binary breast cancer histopathology classification. The framework was evaluated on the BreakHis and IDC datasets using five-fold stratified cross-validation with standard augmentation. It achieved 96.60% accuracy and 98.44% AUC on BreakHis, and 94.52% accuracy and 97.31% AUC on IDC. The model used 6.0 million parameters and 0.60 GFLOPs, indicating that the study considered both classification performance and computational efficiency when developing an architecture for imbalanced histopathological image classification.

[43] investigated the use of pretrained deep learning architectures for automatic seal-integrity inspection in sterile barrier packaging under small and imbalanced manufacturing datasets. The study evaluated multiple CNN architectures on unseen test data and found that model performance generally decreased as the proportion of defective samples became smaller, although some architectures demonstrated greater resilience to class imbalance. ResNet, DenseNet, and EfficientNet variants, particularly B3 and B5, showed strong performance consistency across the experiments, while a U-Net based on VGG16 also performed consistently for segmentation. The study further reported that basic data augmentation had only a marginal effect in some experiments and that overfitting could occur in small manufacturing datasets.

Across these studies, emerging architectures primarily expand the representation-learning and deployment dimensions of imbalance handling. Vision Transformers provide attention-based representations, hybrid CNNs combine complementary feature extractors, transfer learning leverages pretrained representations, ensembles integrate multiple predictors, and knowledge distillation can reduce model size.

However, the reviewed studies differ in dataset size, imbalance severity, application domain, architecture, training protocol, evaluation metric, and computational setting. Therefore, reported performance values cannot be used to establish that a particular emerging architecture consistently outperforms data-level, algorithm-level, or generative approaches.

E. Cross-Theme Synthesis and Research Gaps

The cross-theme synthesis shows that class-imbalance handling can be understood through several intervention levels: modifying the training-data distribution, modifying the learning objective, generating additional minority representations, and integrating imbalance-aware mechanisms with contemporary architectures. These themes represent complementary methodological directions rather than mutually exclusive alternatives.

For RQ1, the reviewed literature identifies several classes of techniques, including SMOTE and ADASYN for resampling, image augmentation for data diversification, class weighting and focal loss for optimization-level

intervention, GAN/VAE/diffusion models for synthetic-data generation, and architecture-integrated mechanisms involving transfer learning, Vision Transformers, ensemble learning, knowledge distillation, and federated learning.

For RQ2, the evidence indicates that reported effectiveness varies substantially across experimental settings. High accuracy values were reported in several studies, but absolute accuracy cannot be used to rank methods across different datasets or application domains. For example, [14] reported 98.63% accuracy and 99.76% AUC in an Alzheimer's MRI setting, whereas [20] reported 82.6% balanced accuracy and 96.4% AUC for diabetic-retinopathy classification. These values cannot be interpreted as a direct comparison because the datasets, class distributions, architectures, and evaluation protocols differ.

A more meaningful interpretation is therefore based on within-study improvement relative to an appropriate baseline and on minority-sensitive metrics. When reported, macro-F1, balanced accuracy, minority-class recall, G-Mean, and AUC provide complementary evidence beyond overall accuracy. The inconsistent reporting of these metrics remains a limitation for direct cross-study comparison.

For RQ3, several factors influence the effectiveness of imbalance-handling strategies. The first is dataset characteristics, including dataset size, class distribution, minority-class diversity, image modality, and domain specificity. The second is imbalance severity, because a technique that performs well under moderate imbalance may behave differently under extreme minority scarcity. The third is architecture, because CNNs, Vision Transformers, ensembles, and lightweight models have different representation and optimization characteristics. The fourth is evaluation protocol, including train-test separation, baseline selection, and metric choice. The fifth is computational cost, particularly for generative, ensemble, federated, and transformer-based approaches.

The comparison of synthetic-data approaches further demonstrates the importance of representation level. Feature-space methods such as SMOTE construct additional representations based on existing feature relationships, whereas image-space augmentation, GANs, and diffusion models modify or generate image representations directly. These approaches should not be treated as interchangeable because they introduce minority information through different mechanisms. Their effectiveness should therefore be interpreted according to representation space, data characteristics, and downstream evaluation rather than the number of synthetic samples alone.

The synthesis also distinguishes generation quality from classification effectiveness. Metrics such as FID, KID, Inception Score, or domain-specific fidelity assessments provide information about synthetic-data quality, whereas accuracy, macro-F1, balanced accuracy, recall, G-Mean, and AUC evaluate downstream classification. A study should therefore not be interpreted as demonstrating classification improvement solely because it successfully generates realistic images. Similarly, improved classification performance does

not automatically demonstrate that generated images are sufficiently diverse or faithful to the target domain.

Computational requirements represent another cross-theme factor. Conventional resampling and augmentation can generally be integrated into existing training pipelines with limited architectural modification. Generative approaches introduce additional model-training or image-generation stages, while ensembles increase the number of models that must be trained or evaluated. Federated learning introduces communication and coordination requirements, and transformer-based or hybrid architectures may require greater computational resources. [17] and [39] illustrate the value of considering inference time or model size together with classification performance.

The evidence therefore does not support the conclusion that one imbalance-handling method is consistently superior across all reviewed settings. This conclusion arises from the heterogeneity of datasets, imbalance severity, architectures, application domains, evaluation protocols, and computational requirements rather than from a ranking of reported accuracy values.

Several research gaps emerge directly from the synthesis. First, the evidence remains strongly dependent on individual datasets and application domains, limiting conclusions regarding cross-dataset and cross-domain generalization. Second, reporting of imbalance severity and class-distribution characteristics is not sufficiently standardized across studies, making it difficult to determine how methods behave under comparable imbalance conditions. Third, evaluation metrics are inconsistent, with overall accuracy still frequently reported without sufficient minority-sensitive measures. Fourth, computational cost, model size, training complexity, and inference requirements are not consistently reported, limiting assessment of practical deployment. Fifth, generative methods require more systematic validation of synthetic-data fidelity, diversity, and domain plausibility in addition to downstream classification performance.

A further gap concerns the limited availability of standardized multi-dataset benchmarks in which several imbalance-handling strategies are evaluated under identical experimental conditions. Without such benchmarks, differences in dataset composition, preprocessing, architecture, and evaluation protocols can obscure whether an observed improvement originates from the imbalance-handling strategy itself or from other components of the experimental pipeline.

For RQ4, these gaps suggest several future research directions. Future studies should prioritize standardized benchmarking across multiple datasets and controlled imbalance ratios, external or cross-dataset validation, consistent reporting of macro-F1, balanced accuracy, minority-class recall, G-Mean, and AUC, and transparent reporting of training and inference costs. For generative approaches, evaluation should combine synthetic-data fidelity and diversity measures with downstream minority-class performance. For emerging architectures, future work should evaluate whether additional architectural complexity

produces measurable benefits relative to its computational and deployment requirements.

Overall, the synthesis indicates that research on imbalanced image classification is moving from isolated data-level balancing toward integrated strategies involving optimization, synthetic generation, transfer learning, advanced architectures, and deployment-aware learning. Nevertheless, the evidence remains heterogeneous. Future progress therefore requires more standardized experimental protocols, stronger cross-dataset validation, minority-sensitive evaluation, transparent computational reporting, and rigorous validation of synthetic data.

IV. CONCLUSION

This Systematic Literature Review synthesized recent approaches for handling class imbalance in deep learning-based image classification based on 31 included studies. The literature was organized into four methodological themes: Traditional Data-Level Approaches, Algorithm-Level and Loss Function-Based Approaches, Generative and Hybrid Approaches, and Emerging Architecture-Integrated Approaches. The synthesis showed that class-imbalance handling has developed from modifying training-data distributions through oversampling and augmentation toward modifying optimization mechanisms, generating minority representations, and integrating imbalance-aware strategies with advanced deep learning architectures.

Traditional data-level approaches such as SMOTE, ADASYN, and image augmentation remained relevant because of their practical implementation and relatively manageable computational requirements. Algorithm-level approaches addressed imbalance by modifying class weighting and loss functions, while generative and hybrid approaches expanded or transformed minority representations using GAN-, VAE-, and diffusion-based methods. Emerging architecture-integrated approaches further combined imbalance-related strategies with transfer learning, Vision Transformers, ensemble architectures, knowledge distillation, and federated learning. These findings indicate that the choice of approach is closely related to the characteristics of the dataset, class distribution, application domain, model architecture, and experimental setting.

The synthesis did not indicate that a single imbalance-handling approach consistently outperformed others across all studies. Reported performance varied according to dataset characteristics, imbalance conditions, architecture, evaluation metrics, validation procedures, and computational requirements. In addition, classification performance could not be interpreted independently of the evaluation strategy, particularly when class distributions were highly imbalanced. Therefore, the effectiveness of an approach should be interpreted within its experimental context rather than based on accuracy alone.

The review also identified several research gaps. Cross-dataset and external validation remained limited, while evaluation practices and the use of minority-sensitive metrics were not fully standardized across studies. Advanced

generative and architecture-integrated approaches also introduced greater computational requirements, and evidence concerning the relationship between synthetic-data quality and subsequent classification improvement remained insufficient in some studies. Future research should therefore emphasize standardized benchmarking, cross-dataset validation, consistent use of minority-sensitive evaluation metrics, transparent reporting of computational requirements, and stronger validation of synthetic-data fidelity and its effect on downstream classification performance. These directions can support more reliable and practically applicable deep learning systems for image classification under imbalanced data conditions.

REFERENCES

- [1] M. Shafiq and Z. Gu, "applied sciences Deep Residual Learning for Image Recognition : A Survey," *Appl. Sci.*, pp. 1–43, 2022.
- [2] J. Plested, M. Phiri, and T. & Gedeon, "Deep transfer learning for image classification: A survey," *Artif. Intell. Rev.*, 2026, doi: 10.1007/s10462-026-11491-z.
- [3] W. Chen, K. Yang, Z. Yu, Y. Shi, and C. L. P. Chen, *A survey on imbalanced learning : latest research , applications and future directions*, vol. 123. 2024. doi: 10.1007/s10462-024-10759-6.
- [4] C. J. Hellín, A. A. Olmedo, and A. Tayebi, "applied sciences Unraveling the Impact of Class Imbalance on Deep-Learning Models for Medical Image Classification," 2024.
- [5] F. Rodríguez-torres and J. F. Martínez-trinidad, "applied sciences An Oversampling Method for Class Imbalance Problems on Large Datasets," 2022.
- [6] M. O. Id, J. Ben Hayfron-acquah, T. Frimpong, and G. Abdul-, "Imbalanced class distribution and performance evaluation metrics : A systematic review of prediction accuracy for determining model performance in healthcare systems," *PLOS Digit. Heal.*, pp. 1–19, 2023, doi: 10.1371/journal.pdig.0000290.
- [7] B. Więckowska and P. & Guzik, "Usmile likelihood evaluation provides robust threshold free assessment of binary classification models for balanced and imbalanced datasets," *Sci. Rep.*, 2026, doi: 10.1038/s41598-026-40545-z.
- [8] D. Holst, K. Moenck, J. Koch, O. Schmedemann, and T. Schüppstuhl, "Transparent Reporting of AI in Systematic Literature Reviews : Development of the PRISMA-trAIce Checklist," *J. Med. Internet Res.*, vol. 4, pp. 1–8, 2025, doi: 10.2196/80247.
- [9] V. Kumari, S. Gupta, and Y. Sharma, "A Multi-Modal Attentive Framework That Can Interpret Text (MMAT)," *IEEE Access*, vol. 13, no. April, pp. 121955–121969, 2025, doi: 10.1109/ACCESS.2025.3586900.
- [10] M. Shahid, M. Ševrović, and M. Olyslagers, "Multi-Scale Spatio-Temporal Feature Aggregation for Road Safety Attributes Classification From Videos," *IEEE Access*, vol. 14, no. May, 2026, doi: 10.1109/ACCESS.2026.3686110.
- [11] H. Bechinia, D. Benmerzoug, and N. Khelifa, "Approach Based Lightweight Custom Convolutional Neural Network and Fine-Tuned MobileNet-V2 for ECG Arrhythmia Signals Classification," *IEEE Access*, vol. 12, no. February, pp. 40827–40841, 2024, doi: 10.1109/ACCESS.2024.3378730.
- [12] S. Jayakumar, V. Sounderajah, P. Normahani, L. Harling, S. R. Markar, and H. Ashra, "Quality assessment standards in artificial intelligence diagnostic accuracy systematic reviews: a meta-research study," pp. 1–13, 2022, doi: 10.1038/s41746-021-00544-y.
- [13] D. Schwabe, K. Becker, and M. Seyferth, "The METRIC-framework for assessing data quality for trustworthy AI in medicine: A systematic review," *npj Digit. Med.* 7, 203, 2024, doi: 10.1038/s41746-024-01196-4.
- [14] M. Muhammad *et al.*, "ADD-Net : An Effective Deep Learning

- Model for Early Detection of Alzheimer Disease in MRI Scans,” *IEEE Access*, vol. 10, no. September, pp. 96930–96951, 2022, doi: 10.1109/ACCESS.2022.3204395.
- [15] A. F. A. Alshamrani, F. Saleh, and Z. Alshomrani, “Optimizing Breast Cancer Mammogram Classification Through a Dual Approach : A Deep Learning Framework Combining ResNet50 , SMOTE , and Fully Connected Layers for Balanced and Imbalanced Data,” *IEEE Access*, vol. 13, no. December 2024, pp. 4815–4826, 2025, doi: 10.1109/ACCESS.2024.3524633.
- [16] A. S. Farhan, M. Khalid, and U. Manzoor, “Combined Oriented Data Augmentation Method for Brain MRI Images,” *IEEE Access*, vol. 13, no. December 2024, pp. 9981–9994, 2025, doi: 10.1109/ACCESS.2025.3526684.
- [17] S. Akhtar, M. Hanif, A. Rashid, and K. Javed, “An Optimized Data and Model Centric Approach for Multi-Class Automated Urine Sediment Classification,” *IEEE Access*, vol. 12, no. May, 2024.
- [18] M. S. Mohammed, A. Al-dhamari, W. Saeed, F. N. Al-aswadi, S. Abdulla, and M. Saleh, “Motion Pattern-Based Scene Classification Using Adaptive Synthetic Oversampling and Fully Connected Deep Neural Network,” *IEEE Access*, vol. 11, no. October, pp. 119659–119675, 2023, doi: 10.1109/ACCESS.2023.3327463.
- [19] V. Hoang, X. Vo, and S. Member, “Categorical Weighting Domination for Imbalanced Classification With Skin Cancer in Intelligent Healthcare Systems,” vol. 11, no. August, 2023.
- [20] W. Nazih and A. O. Aseeri, “Vision Transformer Model for Predicting the Severity of Diabetic Retinopathy in Fundus Photography-Based Retina Images,” *IEEE Access*, vol. 11, no. October, pp. 117546–117561, 2023, doi: 10.1109/ACCESS.2023.3326528.
- [21] F. E. Karahan, “An End-to-End Deep Learning Framework for Crop Type and Phenological Stage Identification,” *IEEE Access*, vol. 14, no. June, pp. 97546–97558, 2026, doi: 10.1109/ACCESS.2026.3707086.
- [22] T. Ali *et al.*, “CAROTIDNet : A Novel Carotid Symptomatic / Asymptomatic Plaque Detection System Using CNN-Based Tangent Optimization Algorithm in B-Mode Ultrasound Images,” *IEEE Access*, vol. 12, no. June, pp. 73970–73979, 2024, doi: 10.1109/ACCESS.2024.3404023.
- [23] V. Yamaçlı, “Battery Usage-Profile Classification via Multimodal Vision Transformer With Feature-Level Fusion and Ensemble Method,” *IEEE Access*, vol. 13, no. September, 2025, doi: 10.1109/ACCESS.2025.3618776.
- [24] X. Qin, F. M. Bui, H. A. H. Nguyen, S. Member, and Z. H. U. Han, “Learning From Limited and Imbalanced Medical Images With Finer Synthetic Images From GANs,” *IEEE Access*, vol. 10, no. September, pp. 91663–91677, 2022, doi: 10.1109/ACCESS.2022.3202560.
- [25] Q. Su, H. Nuzly, A. Hamed, and M. A. Isa, “A GAN-Based Data Augmentation Method for Imbalanced Multi-Class Skin Lesion Classification,” *IEEE Access*, vol. 12, no. February, 2024.
- [26] C. Venkatachalam, “A Dual-Branch Deep Learning Framework Combining Xception and ResNet for Accurate Lung and Colon Cancer Detection,” *IEEE Access*, vol. 13, no. July, pp. 131483–131497, 2025, doi: 10.1109/ACCESS.2025.3589390.
- [27] M. Park, H. W. A. J. Song, and D. Kang, “Imbalanced Classification via Feature Dictionary-Based Minority Oversampling,” *IEEE Access*, vol. 10, pp. 34236–34245, 2022, doi: 10.1109/ACCESS.2022.3161510.
- [28] A. Shafique, M. Siraj, S. A. Alsaif, T. Sadad, and B. Cheng, “Hyperspectral Imaging and Advanced Vision Transformers for Identifying Pure and Pesticide-Coated Apples,” *IEEE Access*, no. April, pp. 66405–66419, 2025.
- [29] G. Mohammad, I. Alam, N. Tasnia, T. Biswas, and S. A. Tanim, “Real-Time Detection of Forest Fires Using FireNet-CNN and Explainable AI Techniques,” *IEEE Access*, vol. 13, no. February, 2025.
- [30] M. Kovac, P. Macsik, V. Kurilova, Z. Fellner, and J. Pavlovicova, “Unconditional GAN-Based Generation of Stage-Specific Fundus Images for Augmenting Diabetic Retinopathy Screening,” *IEEE Access*, vol. 14, no. April, pp. 47800–47815, 2026, doi: 10.1109/ACCESS.2026.3675325.
- [31] A. Asokan and B. R. Manju, “Lesion-Fidelity-Aware Synthetic Data Augmentation for Diabetic Retinopathy Grading Using Conditional GANs,” *IEEE Access*, vol. 14, no. April, pp. 69998–70010, 2026, doi: 10.1109/ACCESS.2026.3690867.
- [32] V. Sreelekshmi, K. Pavithran, and J. J. Nair, “Unleashing the Power of Hierarchical Variational Autoencoder for Predicting Breast Cancer,” *IEEE Access*, vol. 12, pp. 195658–195670, 2024, doi: 10.1109/ACCESS.2024.3518612.
- [33] H. I. Ahmad and M. Ali, “Classification of Psoriasis Using Vision Transformer : A Deep Learning-Based Approach,” *IEEE Access*, vol. 13, no. September, pp. 163885–163895, 2025, doi: 10.1109/ACCESS.2025.3607989.
- [34] S. Chatterjee and J. Gil, “Early Detection of Multiclass Skin Lesions Using Transfer Learning-Based IncepX-Ensemble Model,” *IEEE Access*, vol. 12, no. August, pp. 113677–113693, 2024, doi: 10.1109/ACCESS.2024.3432904.
- [35] R. D. W. I. Prayogo, N. U. R. Hamid, and H. Nambo, “Hybrid CNN-Based Transfer Learning Enhances Brain Tumor Classification on MRI Images,” *IEEE Access*, vol. 13, no. July, 2025.
- [36] E. Silva *et al.*, “Enhancing Diabetic Foot Ulcer Classification Through Fine-Tuned Multilevel CNN,” *IEEE Access*, vol. 13, no. July, 2025, doi: 10.1109/ACCESS.2025.3583111.
- [37] L. Arora, K. K. Yadav, A. Malik, N. Kumar, U. Mamodiya, and R. Jain, “Cross-Modality Transfer Learning for Breast Cancer Detection Using Ultrasound-Pretrained Deep Network on Histopathology Images,” *IEEE Access*, vol. 14, no. May, pp. 76510–76525, 2026, doi: 10.1109/ACCESS.2026.3694268.
- [38] A. V. Sri and M. K. Morampudi, “Privacy-Preserving Federated Learning for Retinal Disease Diagnosis Using Paillier Homomorphic Encryption With Multiple Encryption Keys,” *IEEE Access*, vol. 14, no. January, pp. 26391–26415, 2026, doi: 10.1109/ACCESS.2026.3664444.
- [39] S. M. Vemuri, A. Gundrapally, Z. Xia, J. Kim, and K. K. Choi, “CB-DistillGrad : Class-Balanced Distillation and Gradient Conflict Resolution for Low,” *IEEE Access*, vol. 14, no. December 2025, pp. 3040–3061, 2026, doi: 10.1109/ACCESS.2025.3650071.
- [40] R. F. Rocha, F. H. A. M. Neto, P. H. Ferreira, H. O. Paz, A. K. Suzuki, and F. Louzada, “Addressing Data Imbalance in Medical Imaging : A Simulation Study Using Ensemble Transfer Learning for Skin Cancer Detection,” *IEEE Access*, vol. 14, no. April, pp. 75259–75271, 2026, doi: 10.1109/ACCESS.2026.3692474.
- [41] S. Ahmed, B. Hasan, T. Ahmed, R. K. Sony, and H. Kabir, “Less Is More : Lighter and Faster Deep Neural Architecture for Tomato Leaf Disease Classification,” *IEEE Access*, vol. 10, no. July, pp. 68868–68884, 2022, doi: 10.1109/ACCESS.2022.3187203.
- [42] H. Ali, M. A. Khan, S. Member, G. Barb, and A. Silaghi, “AttFKAN-DEQ : Attention-Enhanced Fourier Kolmogorov-Arnold Networks Using Deep Equilibrium for Breast Cancer Histopathology Classification,” *IEEE Access*, vol. 14, no. May, pp. 91285–91302, 2026, doi: 10.1109/ACCESS.2026.3698647.
- [43] J. Z. Diaz, M. A. L. I. Farooq, and P. Corcoran, “Automatic Inspection of Seal Integrity in Sterile Barrier Packaging : A Deep Learning Approach,” vol. 12, no. January, 2024.