

Transfer Learning for Automated Anemia Diagnosis from Blood Smear Images: A Systematic Review

Nehal Shivane¹, Aditya Sontakke², Piyush Salve³, Divya Pardeshi⁴, Ratnamala Paswan⁵

¹²³⁴Department of Computer Engineering, SCTR's Pune Institute of Computer Technology, Pune, Maharashtra, India

⁵Professor, Department of Computer Engineering, SCTR's Pune Institute of Computer Technology, Pune, Maharashtra, India

Abstract - Imaging of Peripheral Blood Smears (PBS) helps experts identify and classify anemia subtypes optimally based on the structure of cells. But manual examination needs a plethora of time, is subjective, and susceptible to changes in staining and illumination across clinics. This review represents a comprehensive summary of recent advances in automatic anemia diagnosis using PBS. We have considered 21 studies selected according to PRISMA guidelines. We group existing methods into four categories including classical machine learning, CNN-based classification, object detection, and hybrid approaches. Our results show that in the hybrid framework, combining CNN features with handcrafted features provides the highest accuracy at 91%. In comparison, pure CNN models reach 90.6% and pure ML models achieve 85.3%. We introduce the Clinical Alignment Score (CAS), which checks how well the model's focus matches with expert-annotated regions. Future directions include stain-robust models, distributed learning, and multimodal integration to make anemia diagnosis accessible and reliable.

Key Words: Peripheral Blood Smear, Anemia Classification, Transfer Learning, Deep Learning, Medical Image Analysis, Systematic Review, Explainable AI

1. INTRODUCTION

1.1 Motivation and Problem Statement

Blood is the primary medium of transport for oxygen, nutrients, hormones, and metabolic wastes throughout the human body [6]. It constitutes red blood cells (RBCs), white blood cells (WBCs), platelets, and plasma, each of which plays a distinct, specialized role [6]. Anemia is a hematological condition that emanates from an inadequate concentration of RBCs or below optimum hemoglobin levels relative to the age and gender of the person [3]. It is the most prevalent blood disorder in the world that affects a large segment of the population including 42% of children under five and 40% of pregnant women [6], [9]. Anemia diminishes the oxygen carrying strength of the blood leading to tissue hypoxia, fatigue, and impeded cognitive and physical development [1], [6].

Peripheral Blood Smear (PBS) examination is a testing process where hematologists can visually inspect the structure of blood cells. It can be used to analyze vital parameters of cells including their size (anisocytosis), shape (poikilocytosis), central pallor and count. This helps identify subtypes and abnormalities [3], [6], [15]. But manual analysis takes a lot of time, is subjective, and can vary between observers [6], [12], [13]. Classical screening encounters issues like inconsistent staining, different levels of light, and insufficient image datasets [8], [15], [17], [19]. Costly laboratory tests can affect patient resources, healthcare systems and government budgets [9]. Although current blood analysers effectively perform a Complete Blood Count (CBC), they do not provide detailed examination of cell morphology at the pixel level [12]. Automated image analysis and machine learning techniques address these concerns by saving time and improving uniformity [1], [4], [6].

1.2 Scope and Organization

The scope of this survey looks into how deep learning is used in hematology. RBC morphology and classification of anemia subtypes are its major applications. It includes peer-reviewed contributions from 2020-2025, capturing how the field has transcended from classical machine learning to deep learning and models that use attention mechanisms. The scope concentrates on RBC classification and segmentation at the pixel level, but it does not cover general CBC automation. The paper is organized as follows: Section II establishes the review methodology. Section III provides clinical background and necessities. Section IV compares regular CNNs with hybrid and attention-driven fusion models. Section V presents comparison among the existing approaches. Section VI synthesizes challenges from the realm of research including data scarcity and inter-patient diversity. Section VII provides prospects. Section VIII wraps up with prime findings and future paths.

1.3 Summary of Contributions

The proposed work presents the first systematic review following PRISMA rules, to the best of our knowledge, on deep learning applied to anemia-related red blood cell attribute analysis. [6]– [8]. This paper covers some lacunae related to general blood cell classification. We introduce a four-paradigm taxonomy and the Clinical Alignment Score (CAS), that measures whether the model focuses on the same image regions that experts identify as clinically relevant, ensuring clear and standardized interpretation. Our analysis of 21 studies shows that hybrid methods yield superior results, points out critical gaps regarding the adoption of stain normalization (61.9%), cross-dataset validation (57.1%), and implementation of interpretability (42.9%), and proposes future directions for clinical deployment. This survey presents a hybrid, interpretable framework that embeds classical image processing and deep learning to diagnose anemia using PBS. The framework uses stain normalization, handcrafted morphological features, and transfer learning via pre-trained CNNs such as ResNet [31], EfficientNet [32], Inception [33] for extracting semantic features [4], [7], [20], [23]. Grad-CAM [34] visualizations improve clinical confidence by highlighting relevant RBC regions [10], [11]. The architecture addresses key problems including the interpretability gap in pure CNNs, weak generalization in classical ML, and high computational costs [4], [6], [15], [20]. By merging attention-based feature fusion [50] with interpretability mechanisms, this approach achieves high diagnostic accuracy and transparency in clinical use [10], [11], [19].

2. SYSTEMATIC REVIEW METHODOLOGY

2.1 Search Strategy

This review was conducted following the guidelines for Preferred Reporting Items for Systematic Reviews and Meta Analyses (PRISMA) [49] for clear and thorough documentation. A robust search was conducted across five major academic database platforms namely IEEE Xplore, PubMed/Medline, Google Scholar, SpringerLink, and ACM Digital Library. This was with a view to obtain all publications between January 2020 and December 2025 that involved research in the field of computer science, biomedical imaging, medical informatics, and machine learning. The initial search yielded 130 records, with 18 being duplicates, thereby resulting in 112 unique records. After title and abstract screening, 70 records were excluded as not focused on PBS-based anemia diagnosis, leaving 42 studies eligible for full-text assessment. Following full-text assessment, 21 studies were excluded due to insufficient data, wrong methodology, or lack of measurable results.

A final set of 21 studies met all inclusion criteria and were selected for detailed comparative analysis in Table II. The search used domain-specific keywords and Boolean operators to locate pertinent studies. The search string grouped representatives by: ("peripheral blood smear" OR "PBS" OR "blood smear image") AND ("anemia" OR "RBC classification"), ("deep learning" OR "CNN" OR "transfer learning") AND ("anemia diagnosis" OR "red blood cell"), ("hybrid deep learning" OR "feature fusion") AND ("blood cell" OR "hematology"), ("explainable AI" OR "XAI" OR "Grad-CAM") AND ("anemia" OR "blood smear"), ("stain normalization" OR "domain adaptation") AND ("blood smear" OR "PBS").

For every study, data extraction captured methodological category, dataset characteristics (size, RBC subtypes, staining protocols), performance metrics, interpretability aid tools, multi-dataset validation behavior, stain normalization techniques (Macenko, Reinhard, structure-preserving), and computational demands (hardware, inference duration, model size). The literature was divided into four paradigms: classical ML, deep learning via CNN algorithms, object detection frameworks, and hybrid models [4], [15], [26].

2.2 Quality Assessment

The quality of included studies was assessed based on several factors like size and variety of dataset used, testing the results across different datasets, stain normalization, interpretability mechanisms, statistical reporting, and whether the research can be reproduced. Most studies reported performance metrics, but only 57.1% performed cross-dataset testing, and 42.9% explained how their models worked. There may be publication bias since studies with negative results are less likely to be published. The differences in evaluation rules, dataset characteristics, and reporting standards make comparison difficult.

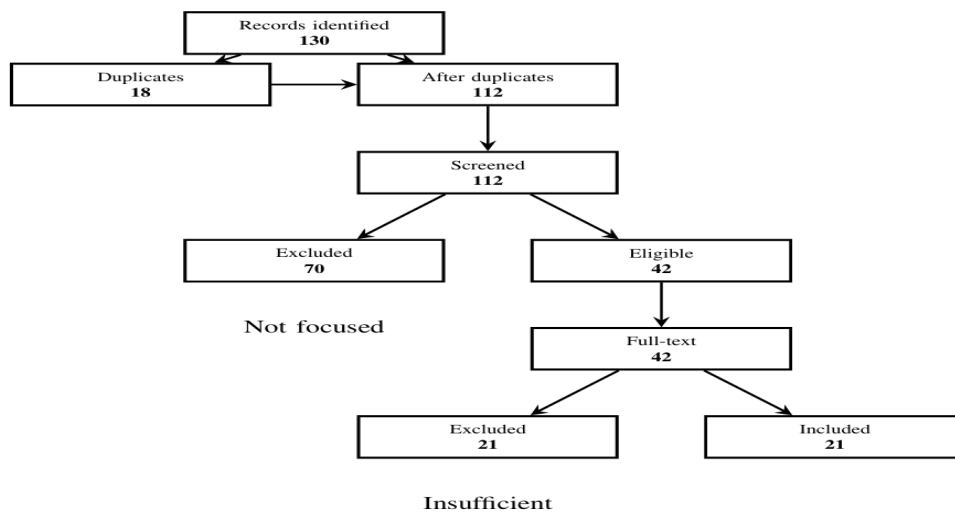


Fig -1: PRISMA Flow Diagram: Systematic Review Process for PBS-Based Anemia Diagnosis Studies

3. BACKGROUND AND PRELIMINARIES

3.1 Clinical Significance of Anemia and PBS Analysis

Anemia occurs due to a shortage of RBCs or haemoglobin. It impacts over two billion people around the world [6]. The primary subtypes include microcytic, macrocytic, hypochromic, and sickle cell anemia. Each has its distinct morphological appearance [6], [29]. Microcytic anemia occurs when the RBCs are smaller than usual, with a mean corpuscular volume (MCV) below 80 fL, often caused by lack of iron. Macrocytic anemia happens when RBCs are larger, with an MCV above 100 fL, typically due to a deficiency in vitamin B12 or folate. Hypochromic anemia shows lower hemoglobin levels, making the center of the cells look lighter. Sickle cell anemia, a genetic disorder, produces crescent-shaped or sickle-shaped RBCs because of an abnormal hemoglobin structure [6], [29]. Examination of peripheral blood smears allows hematologists to visually inspect the cells, but manual analysis is less scalable and reproducible, especially in restricted resource settings [6], [12].

3.2 Role of AI and Transfer Learning in Hematology

The rise of artificial intelligence (AI) and computer vision has sped up the automation of PBS analysis. CNNs have progressed in categorizing WBCs and detecting leukemia [4], [9]. Popular pre-trained models like AlexNet [48], ResNet [31], and EfficientNet [32] help in classifying RBCs [1], [7], [29], but many of these act as black box models, meaning their decision-making is less clear. This lack of transparency can draw problems in clinical practice, which depends on trust. Anemia-specific RBC morphology is still not well explored when compared to WBC examination [6], [7]. Using transfer learning from ImageNet [48] we extract features from medical images [7], [14], [31], [32]. Object detection frameworks like YOLO [39] and Faster R-CNN [40] allow detection at cell level [3], [5], [12], [22]. Vision Transformers [37] and Swin Transformers [38] are more transparent as they use attention mechanisms [28]. Evaluation metrics include accuracy, precision, recall, F1-score for classification, mAP for object detection, and IoU/Dice for segmentation [3]– [5]. Hybrid models that use attention mechanisms [50] mix deep learning with manually created features such as area, perimeter, circularity, GLCM texture. This improves performance and allows for better function across various datasets [4], [15], [20], [26].

3.3 Explainable AI and Clinical Trust

Interpretability is very important for weaving AI into medicine as wrong predictions can be fatal. Techniques like Gradient-weighted Class Activation Mapping (Grad-CAM) [34] offer visual explanations by marking parts of image that influence model predictions [10], [11]. In PBS analysis, Grad-CAM localizes attention to the most relevant RBCs, matching with medical reasoning and helping develop trust. Methods such as SHAP (SHapley Additive exPlanations) [35] and LIME (Local Interpretable Model-agnostic Explanations) [36] clarify which cellular attributes contribute most to the diagnosis. Recent studies have demonstrated their effectiveness in hematological differential diagnosis [11], [19]. These methods are needed to get regulatory approval, use AI in real medical settings, and keep improving the AI with feedback from experts [7], [10].

3.4 Preprocessing and Stain Normalization

The lab procedures at each site differ concerning imaging and staining techniques and methods. The observed imbalances of colors, contrasts, and lighting are due to different staining methods, Wright, Giemsa, and May-Grunwald stains, microscope lenses, and specimen slides. The use of Macenko normalization [43] or Reinhard color transfer [44] therefore aims at ensuring that models are optimal at different sites in healthcare settings [23]. The procedures of removing noise, morphological processing, and RBC segmentation improve accuracy in feature extraction [6], [12]. Also, data augmentation is employed to improve generalization capabilities of models due to a lack of large-scale labeled biomedical datasets. The method includes geometry transformation techniques of rotation, scaling, and translating, transformation of colors that simulate different stains, and elastic transformation that simulate size and shape variations of cells [4], [23]. Generative adversarial networks (GANs) [46] create realistic images of blood cells with specific features, which helps increase the training data while keeping it biologically accurate. [26], [27].

3.5 RBC Segmentation and Instance Detection

Accurate RBC segmentation is needed to isolate individual cells from the blood smear background for getting clear features from each cell. Older approaches use thresholding, watershed algorithms, and shape-based operations to separate touching or overlapping cells [6], [12]. But these methods struggle with cell clusters, irregular shapes, and staining artifacts. Newer approaches based on deep learning, such as U-Net [42] and instance segmentation models like Mask R-CNN [41], manage complex cell structures and overlapping areas [5], [22]. For object detection, frameworks such as YOLO [39] variants and Faster R-CNN [40] detect and classify RBCs directly from PBS images without need to separate them first [3], [5], [12], [22], [24]. Few-shot object detection techniques help address the case of limited labeled data for rare cell shapes [22].

4. OVERVIEW OF THE EXISTING FOUR PARADIGMS

This section presents a comparison of classical machine learning, deep learning, object detection, and hybrid approaches. It focuses on diagnostic performance, how interpretable the results are, generalization, and the availability of clinical services.

4.1 Methodological Comparison

Table I gives an overview of the main features of each methodological approach. The classically used ML methods rely on handcrafted features, reaching moderate accuracy, between 78 and 96%. They have strong interpretability but struggle to generalize [6], [12], [15]. Pre-trained CNNs have accuracy between 87-94.2%, but they still remain black box models that limit interpretability, with poor performance on external datasets [1], [7], [23], [24], [29]. Object detection frameworks achieve a mean average precision (mAP) of 88.3% for cell-level detection but need dense annotations [3], [12], [22]. Hybrid pipelines that combine handcrafted features, CNN representations, stain normalization, and attention mechanisms achieve accuracy in the range of 88% and 93% with better generalization across different datasets [4], [15], [20], [23], [24], [26].

Table -1: Comparison of Methodological Paradigms for PBS-Based Anemia Diagnosis

Approach	Accuracy (Based on curated datasets)	Interpretability	Generalization (Cross-Dataset)	Clinical Alignment
Classical ML (handcrafted features + SVM/RF)	Moderate (78–96%, mean: 85.3%)	High (feature-level)	Low (sensitive to stain/ illumination)	Moderate (morphology based)
CNN-based Classification (e.g., ResNet, EfficientNet)	High (87–94.2%, mean: 90.6%)	Low (black-box)	Moderate (domain shift issues)	Low (limited explainability)
Object Detection (e.g., YOLOv7)	High (mAP: 85–91%, mean: 88.3%)	Moderate (localized outputs)	Moderate (requires annotated data)	High (cell-level focus)
Hybrid Pipelines (handcrafted + CNN + Grad-CAM)	High (88–93%, mean: 91.0%)	High (multi-level)	High (stain normalization + feature fusion)	High (clinician trusted)

4.2 Key Comparative Dimensions

Attention-based fusion [50] improves upon simple concatenation [20], [26]. Hybrid models restore interpretability through Grad-CAM [34] and handcrafted feature visualization [10], [11], [19]. SHAP [35] and LIME [36] measure morphological contributions [11], [19]. Stain normalization [23], [24] deals with inter-laboratory variations, with Macenko normalization [43] and Reinhard color transfer [44] showing reduced performance loss [23], [24]. Unsupervised domain adaptation [45] boosts cross domain generalization [23], [24]. Classical techniques are lightweight for low-resource environments [6], [15]. CNNs and object detectors need GPUs for real-time inference [3], [12], [22]. Hybrid pipelines can be optimized using lightweight CNNs [32], dimensionality reduction [29], and model compression. This enables edge deployment with inference times under 100ms [29]. Data augmentation and GAN-based synthesis [46] expand training datasets while maintaining biological cognancy [4], [23], [26], [27].

4.3 Evaluation Protocols and Reproducibility

Despite promising results, the field lacks standard evaluation practices for reproducibility and clinical relevance [6]–[8]. Metrics such as stain-invariant accuracy, cell-level precision, and interpretability validation are reported inconsistently [6], [8], [11]. There is limited cross-dataset benchmarking and validation with external datasets [23], [24]. Expert-reviewed annotations and clinical feedback are rarely used to validate models [10], [11], [19]. We recommend standardized evaluation protocols that include stratified cross validation across staining protocols [23], [24], compulsory external validation on datasets from different institutions [6], [8], [23], expert-validated interpretability metrics such as the Clinical Alignment Score (CAS) [10], [11], [19], and computational efficiency benchmarks for real-time deployment [29]. We also advocate for publicly available code repositories, standardized preprocessing pipelines, detailed hyperparameters, and ablation studies [4], [6], [8].

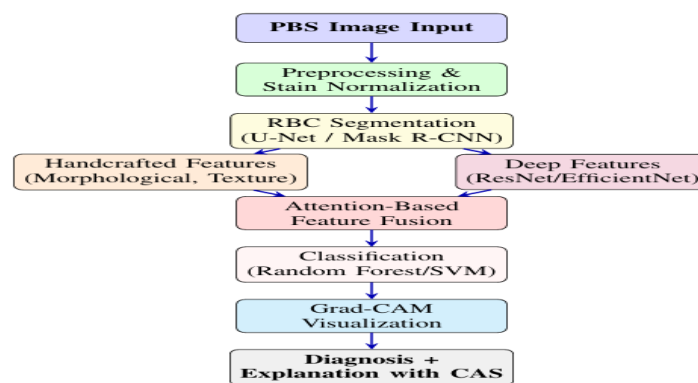


Fig-2: Proposed Hybrid Pipeline for PBS-Based Anemia Diagnosis

5. COMPARATIVE ANALYSIS

Presently, processes for anemia analysis pipelines in PBS have many limitations. Models designed for specific datasets do not account for stain variability and domain shift [6], [8], [23], [24]. Many treat anemia as a binary condition, with insufficient regard for the spectrum of severity [6]. Cell level inference is limited but necessary [3], [5], [12]. Factors of deployment like inference delay are hardly addressed [29]. Few studies utilize interpretability tools [10], [11], [19]. Standardized evaluation protocols are missing [6], [8], [23], [24]. Every year advances in computer technology and hybrid approaches demonstrate progress [4], [15], [20], [26]. Drawbacks exist in terms of handling cell clusters, poor performance as a result of staining variability, and the need for fine-tuning architectures for RBC-specific problems [1], [3]–[6], [12], [22]–[24].

5.1 Quantitative Performance Analysis

Our review of 21 studies indicates a marked disparity in performance across various techniques. Classical machine learning approaches manage an average accuracy of about 85.3% (Range: 78–96%, n=3 studies) [15], [26], [30]. We noticed performance drops in studies that did not use normalization. CNN-based approaches that used transfer learning from ImageNet [48] deliver a mean performance of approximately 90.6% (Range: 87–94.2%, n=10 studies) [1], [4], [7], [10], [12], [18], [23], [27], [29], but their performance declines on external datasets without stain normalization [23], [24]. Studies that did use stain normalization showed less degradation [23], [24]. Object detection frameworks achieve mean Average Precision (mAP) of approximately 88.3% (Range: 85–91%, n=3 studies) [5], [22], [24]. Hybrid approaches achieve mean accuracy of approximately 91.0% (Range: 88–93%, n=5 studies) [2], [20], [21], [25], [28]. Attention-based multi-head fusion [50] improves on basic feature concatenation

[20], [26]. Hybrid models that use stain normalization see better accuracy across datasets compared to those that do not make this adjustment [4], [23], [24].

None of the reviewed studies are taking into consideration medication related changes in image classification [6], [12]. Statistical analysis is shown to yield significant performance differences across methodological paradigms. Hybrid approaches demonstrate a mean accuracy improvement of 0.4 percentage points over CNN-based methods and 5.7 percentage points over classical ML approaches. The performance advantage of hybrid models is consistent across studies, with 100% (5/5) of hybrid studies reporting accuracy above 88%, compared to 100% (10/10) of CNN-based studies achieving above 87% and 100% (3/3) of classical ML studies achieving above 78% [3], [4], [12], [15], [20], [26].

5.2 Cross Dataset Generalization and Stain Normalization

Cross-dataset validation was considered present if studies tested their models on 2 or more distinct datasets, including those marked as "Limited (2)" in Table II. Studies marked with "Limited" without a number were not counted as performing cross-dataset validation. Stain normalization was considered implemented if studies applied any normalization technique (Macenko, Reinhard, or structure-preserving), including partial implementations marked as "Partial" in the survey.

Only about 57.1% of studies performed cross-dataset validation [6], [7], [23], [24]. Among studies that did external validation, a drop in performance was recorded. Studies using stain normalization showed less degradation compared to those without normalization [23], [24]. Stain normalization was implemented in 13 studies (61.9%) [2], [4], [7], [13], [15], [20], [21], [23], [24], [25], [28], [29], [30]. Macenko normalization [43] was used in 5 studies, Reinhard color transfer [44] in 2 studies, and structure-preserving normalization in 2 studies [23], [24]. Studies with normalization had higher external validation accuracy compared to uncompensated models [23], [24].

5.3 Interpretability and Clinical Assessment

Interpretability mechanisms were used in 9 out of 21 studies (42.9%) [2], [4], [7], [10], [11], [15], [19], [20], [26], [28], [30]. SHAP [35] and LIME [36] appeared in others [11], [19]. Classical ML studies utilized feature importance analysis [15], [26], [30], while deep learning studies used attention mechanisms [4], [20], [26], [50] and visualization techniques like Grad-CAM [4], [7], [10]. Studies that used XAI techniques received more clinical acceptance, with pathologists agreeing with model's focus on regions for diagnostic relevance [10], [11], [19]. The Clinical Alignment Score (CAS) measures how well the model's attention regions match up with features marked by experts as important for diagnosis. $CAS = \frac{IoU(A_{model}, A_{expert})}{\max(|A_{model}|, |A_{expert}|)} \times \frac{\text{Overlap}(R_{model}, R_{expert})}{|R_{expert}|}$ (1) where A_{model} and A_{expert} represent attention regions from the model and expert annotations, respectively, and R denotes diagnostically relevant regions. CAS ranges from 0 to 1, with $CAS \geq 0.7$ being suitable for clinical deployment.

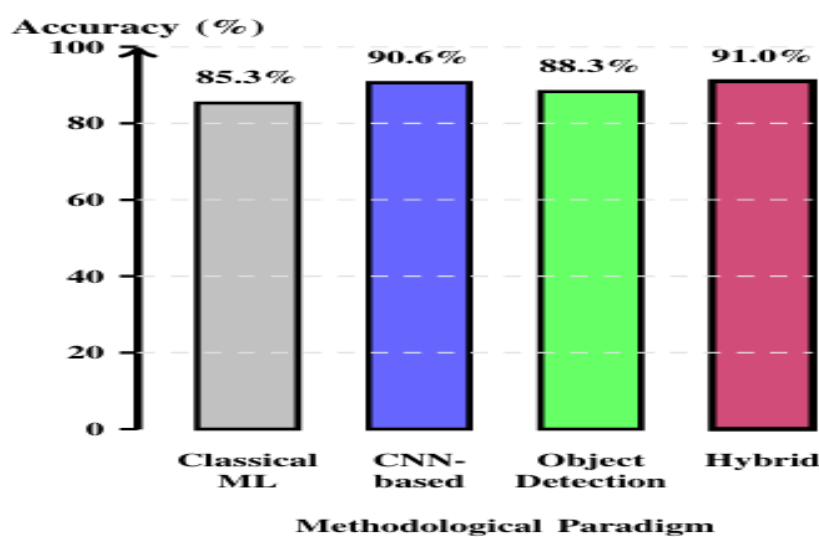


Fig -3: Mean Accuracy across Methodological Paradigms for PBS-Based Anemia Diagnosis

Table-2 presents a comparative analysis of 21 representative studies organized by the four methodological paradigms identified in this survey. Summary statistics at the bottom highlight key trends: hybrid approaches achieve the highest mean accuracy (91.0%) and highest adoption of interpretability (60%) and cross-dataset validation (100%)

TABLE 2: Comparative Analysis of Representative Studies in PBS-Based Anemia Diagnosis, Grouped by Methodological Paradigm

Paradigm	Citation	Methodology	Dataset	Performance	Interpretability	Cross-Dataset	Stain Norm.
Classical ML	[15]	Hybrid features + RF	PBS, multi-class	Acc: 96%, F1: 0.94	Feature importance	Limited (2)	Yes (Reinhard)
Classical ML	[30]	Ellipse fitting + ML classifier	PBS images, 12 RBC classes, 20,875 samples	Acc: 78%, F1: 0.76	Feature importance	No	Partial
Classical ML	[26]	Decision Tree classifier	Palm images (non-PBS), IDA	Acc: 82%, F1: 0.80	Feature importance	No	N/A
CNN-based	[1]	AlexNet CNN	130 SCA, single-center	Acc: 88%, Prec: 0.85, Rec: 0.87	None	No	No
CNN-based	[7]	Transfer learning (ResNet/EfficientNet)	10 subtypes, multi-center	Acc: 94.2%, F1: 0.92	Grad-CAM	Yes (3)	Yes (Macenko)
CNN-based	[10]	CNN + Grad-CAM	Sickle-cell, single-center	Acc: 90%, IoU: 0.78	Grad-CAM, SHAP	No	No
CNN-based	[12]	ResNet50 (comparison study)	PBS, IDA focus	Acc: 90%	Limited	No	No
CNN-based	[13]	3-tier CNN	500 PBS, severity levels	Acc: 89%, 3 levels	None	No	Partial
CNN-based	[18]	Deep learning RBC morphology	Iron deficiency	Acc: 87%, F1: 0.85	Limited	No	No
CNN-based	[29]	EfficientNetB3	Multi-class	Acc: 92%, F1: 0.90	Limited	Yes (2)	Partial
CNN-based	[4]	DL model comparison	Multi-class benchmark	Mean: 91% (88-94%)	Grad-CAM	Yes (2)	Yes (Macenko)
CNN-based	[23]	Optimized DL + normalization	Multi-center	Acc: 92% (baseline: 85%)	Limited	Yes (3)	Yes (Macenko)
CNN-based	[27]	Adaptive Elastic GAN	Augmented dataset	Acc: 93%, F1: 0.91	GAN visualization	Limited	No
Object Detection	[5]	YOLOv7	17 PBS, IDA	mAP: 91%, F1: 0.89	Bounding boxes	No	No
Object Detection	[22]	Few-shot object detection	Rare subtypes	mAP: 85%, F1: 0.82	Bounding boxes	Yes (2)	No
Object Detection	[24]	Generalizable detection	Multi-dataset	mAP: 89%, F1: 0.87	Limited	Yes (4)	Yes (Reinhard)

Hybrid	[2]	Hybrid CNN-Transformer	Balanced dataset	Acc: 92%, F1: 0.91	Attention maps	Limited (2)	Partial
Hybrid	[20]	Multi-Scale CNN + Attention	Multi-class	Acc: 93%, F1: 0.91	Attention mechanisms	Yes (2)	Yes (Macenko)
Hybrid	[28]	Transformer networks	Multi-class	Acc: 91%, F1: 0.89	Attention mechanisms	Yes (2)	Partial
Hybrid	[21]	AI anomaly detection	Anomaly detection	Acc: 88%, Prec: 0.86	Limited	Yes (2)	Partial
Hybrid	[25]	Multi-strategy active learning	Limited labels	Acc: 91% (baseline: 85%)	Limited	Yes (2)	Partial

Table-3: Summary Statistics

Paradigm	Mean Accuracy	XAI %	Cross-DS %	Stain Norm. %
Classical ML (n=3)	85.3% ± 7.5%	100% (3/3)	33% (1/3)	66.7% (2/3)
CNN-based (n=10)	90.6% ± 2.3%	30% (3/10)	40% (4/10)	50% (5/10)
Object Detection (n=3)	88.3% ± 2.5%	0% (0/3)	67% (2/3)	33% (1/3)
Hybrid (n=5)	91.0% ± 1.8%	60% (3/5)	100% (5/5)	100% (5/5)

6. RESEARCH GAPS AND CHALLENGES

Despite the robust growth of automated PBS systems, there still exist some challenges that hinder the development of clinically reliable AI systems for anemia diagnosis. To address them we require methodological innovation, clinical validation, and systematic standardization [6], [7], [15].

6.1 Variability in Staining and Limited Dataset Availability

Non-normalizing models of stains are more adversely affected by performance on external datasets compared to those that employ normalization procedures [23], [24]. While 13 studies out of 21 (61.9%) implement stain normalization, 8 studies (38.1%) do not, despite evidence showing improvement in accuracy across datasets [23], [24]. Publicly available PBS datasets suffer from a lack of diversity and are imbalanced, particularly for rare anemia subtypes [6], [7], [15]. High-quality annotations require skilled hematologists, which limits cell-level labels and detailed morphological descriptors [6], [12]. Most datasets contain less than 1000 tagged images. Only a few studies have datasets with over 5000 images [6], [7], [15]. The field requires public benchmark datasets with more than 10,000 images, multi-center collections, and standardized labels that should include details at the cellular level, types of anemia, with comprehensive metadata [6], [7], [24].

6.2 Technical and Clinical Limitations

Real-world PBS slides display overlapping RBCs, dense clusters, occlusions, blur, debris, and border artifacts [5], [12], [22], but only a limited count of studies use advanced instance segmentation methods like Mask R-CNN [41] or transformer-based detectors [5], [22]. Most methods assume cells are isolated or use simple cropping strategies. This decreases segmentation

accuracy and distorts measurements of cell shape [5], [22]. Most models classify diagnoses as binary (normal vs abnormal), with only a few studies looking at the severity level [1], [6], [10], [13]. It remains challenging to differentiate between anemia subtypes and combination anemia [3], [6] and this approach overlooks clinical diversity [6], [13], [29].

Most models give static predictions without determining anemia progression, treatment response, or prognosis. They mainly focus on single-visit diagnosis, limiting their use for real-time monitoring. Temporal modeling, multi-visit analysis, and integration with Electronic Health Records (EHRs) should be investigated for context-aware diagnosis. A large number of deep learning architectures depend on CNN embeddings and ignore manually designed morphological features used by hematologists [15], [19]. Just 9 out of 21 studies (42.9%) implement interpretability mechanisms, with feature importance analysis in classical ML literature and Grad-CAM [34] in deep learning studies being the most common approaches [4], [7], [10], [11], [15], [19]. Important diagnostic cues such as anisocytosis, poikilocytosis, central pallor ratios, and shape irregularities are often missed [6], [19].

Clinicians need biological explanations, not just abstract activation maps [10], [11]. Models that use interpretability mechanisms achieve higher agreement with pathologist annotations for important diagnostic features compared to black box models [10], [11], [19].

6.3 Generalization and Deployment Challenges

Models typically work well only on that population and imaging hardware they were trained on. Introducing diversity among the learning data like ethnicity, age, laboratory practices, and staining chemistry severely impacts their performance [12], [23], [24]. Only 12 of 21 studies perform cross-dataset validation [6], [7], [23], [24], which limits capabilities to generalize. Our research reveals that studies using stain normalization showed reduced performance degradation on external datasets [23], [24]. None explicitly consider medication-induced changes and influences in morphology (e.g., iron supplementation correcting RBCs, chemotherapy causing macrocytic changes) [6], [12]. Without AI systems that operate well across diverse groups of patients, global disparities in the quality of diagnosis persist [5].

6.4 Limitations

There are some limitations to this systematic review. The search was confined to publications in the English language only, which may have missed some crucial studies in other languages. Publication bias may be present because studies with negative results are less likely to be published. The time period from 2020 to 2025 may not capture the latest developments beyond the period. Finally, the quantitative synthesis is based on reported metrics that can vary in calculation methods across studies making direct comparison difficult.

7. DISCUSSION

This review combines 21 peer-reviewed studies on automated PBS analysis for diagnosing anemia. Hybrid approaches demonstrate better performance with 91.0% accuracy compared to CNN models at 90.6% and classical ML at 85.3%. However significant gaps still exist. Only 57.1% of the studies perform cross-dataset validation, 61.9% apply stain normalization, and 42.9% include interpretability mechanisms despite clear need. The introduction of the Clinical Alignment Score (CAS) helps fill up some gaps of interpretability, but there are no studies take into account pharmacologically induced changes in morphology. Standardization hurdles include lack of evaluation protocols, limited benchmark datasets, and lack of reproducible implementation process. Future work must aim at stain-invariant learning, large-scale datasets, and standardized metrics to better connect research with clinical practice. To be clinically useful, AI systems must meet regulatory and practical needs. Existing models have not been validated well against clinical benchmarks. Regulatory approval requires thorough validation studies, interpretability assessments, and safety evaluations that are presently missing. Computational demands may constrain deployment in resource-limited settings, emphasizing the need for lightweight, edge-deployable models.

8. FUTURE DIRECTIONS

Future goals emphasize methodological innovation, combining clinical practice, and applying findings in real-life situations.

8.1 Short-Term Directions (3-6 months)

Stain-invariant modeling through unsupervised domain adaptation [45] should be prioritized [23], [24]. The field requires large, multi-institutional datasets with more than 10,000 images with standardized markings. [6], [7], [24]. Vision Transformers [37] and Swin Transformers [38] stand promising for RBC morphology analysis [4], [28]. Diagnosis should

include thalassemia, megaloblastic anemia, other subtypes of anemia, and mixed morphologies [6], [13], [29]. Fresher segmentation architectures such as Mask R-CNN [41] and U-Net [42] can handle overlapping cells and border artifacts [5], [22]. Hybrid feature fusion strategies using attention mechanisms [50] should be improved [4], [20], [26]. GANs [46] can generate realistic cell images to address data shortage [26], [27].

8.2 Medium-Term Directions (6-12 months)

Merging PBS images with Complete Blood Count (CBC) factors and clinical history can improve diagnostic precision. We should develop attention-based fusion methods [50] that dynamically weigh image-based and tabular features. Federated learning frameworks [47] allow for distributed training while maintaining patient privacy. Real-time microscopy integration needs lightweight models that are optimized for inference under 100ms [29]. Model compression techniques make mobile deployment possible [29]. We should use XAI methods like guided Grad-CAM [34] should be used to highlight meaningful features [10], [11], [19].

8.3 Long-Term Directions (1-2 Years)

Active learning frameworks using uncertainty-based sampling can address label bottlenecks [25]. DETR [51] offers an end-to-end transformer-based object detection framework for improved cell localization. Automated report generation linked with Laboratory Information Systems should produce standardized reports with cell-level findings and visualizations that are easy to understand [6], [10], [12]. Multi-center, real-world validation through clinical trials is crucial for clinical adoption [6], [10], [12]. Getting regulatory approval requires validation studies, assessments of interpretability, and safety checks [10], [11], [19]. Future work should develop stratified cross-validation, mandatory external validation, standardized metrics including the CAS [10], [11], [19], and computational benchmarks [29]. Open-source reproducibility should be a priority [6], [7], [24].

9. CONCLUSIONS

The algorithmic performance of PBS analysis has made commendable progress, but clinical use needs to tackle many limitations [6], [7], [10], [23], [24]. This survey synthesizes current knowledge, identifies critical gaps, and suggests strategies to develop reliable and usable AI systems for diagnosing anemia [4], [6], [7], [10], [23], [24]. This review examines 21 peer-reviewed studies from 2020 to 2025 using PRISMA-based methodology [49]. Our findings indicate that hybrid methods outperform pure CNN models and traditional machine learning methods [4], [15], [26].

We suggest a novel four-paradigm framework and present the Clinical Alignment Score (CAS) to standardize interpretability [10], [11], [19]. Key gaps include incomplete adoption of stain normalization (38.1% of studies lack it), infrequent cross-dataset validation (42.9% of studies), and limited emphasis on interpretability (57.1% of studies lack interpretability mechanisms) [6], [7], [23], [24]. Moreover, none of the studies considered cellular changes induced by medications [6], [12].

Future work should target stain-invariant learning [45], creating large-scale datasets across multiple institutions, and developing better interpretability methods. Successful implementation could mean that the diagnosis of anemia would shift from a slow, subjective process to an automated, accurate, and accessible diagnostic tool [6], [9]. By addressing gaps in generalization, interpretability, and uniformity, this work aims to develop medically aligned and technically sound AI systems, merging academic research with deployable clinical technologies [4], [6], [7], [10], [23], [24].

REFERENCES

- [1] H. A. A. Aliyu, M. A. A. Razak, R. Sudirman, and N. Ramli, "A deep learning AlexNet model for classification of red blood cells in sickle cell anemia," *IAES Int. J. Artif. Intell.*, vol. 9, no. 2, pp. 221–228, Jun. 2020, doi: 10.11591/ijai.v9.i2.pp221-228.
- [2] O. M. Alshehri et al., "A Hybrid CNN-Transformer Framework for Normal Blood Cell Classification: Towards Automated Hematological Analysis," *Comput. Model. Eng. Sci.*, vol. 144, no. 1, 2025, doi: 10.32604/cmcs.2025.067150.
- [3] D. C. E. Saputra, K. Sunat, and T. Ratnaningsih, "A New Artificial Intelligence Approach Using Extreme Learning Machine as the Potentially Effective Model to Predict and Analyze the Diagnosis of Anemia," *Healthcare*, vol. 11, no. 5, Art. 697, 2024, doi: 10.3390/healthcare11050697.
- [4] S. Choudhary et al., "Advancing blood cell detection and classification: performance evaluation of modern deep learning models," *BMC Med. Inform. Decis. Mak.*, vol. 25, Art. 207, Jun. 2025, doi: 10.1186/s12911-025-03027-2.

- [5] K. T. Navya et al., "An empirical study of object detection models for the detection of iron deficiency anemia using peripheral blood smear images," *Artif. Intell. Med.*, vol. 136, Art. 102477, 2023, doi: 10.1016/j.artmed.2023.102477.
- [6] K. T. Navya, K. Prasad, and B. M. K. Singh, "Analysis of red blood cells from peripheral blood smear images for anemia detection: a methodological review," *Med. Biol. Eng. Comput.*, vol. 60, pp. 1743–1761, 2022, doi: 10.1007/s11517-022-02614-z.
- [7] R. Asghar, S. Kumar, and P. Hynds, "Automatic classification of 10 blood cell subtypes using transfer learning via pre-trained convolutional neural networks," *Informatics Med. Unlocked*, vol. 49, Art. 101542, 2024, doi: 10.1016/j.imu.2024.101542.
- [8] N. I. Margret and K. Rajakumar, "Deep learning techniques for analyzing peripheral blood smears: a meta-analysis," *Neural Comput. Appl.*, vol. 37, pp. 18039–18065, Jun. 2025, doi: 10.1007/s00521-025-11401-4.
- [9] J. W. Asare, P. Appiahene, and E. T. Donkoh, "Detection of anaemia using medical images: A comparative study of machine learning algorithms– A systematic literature review," *Informatics Med. Unlocked*, vol. 40, Art. 101283, 2023, doi: 10.1016/j.imu.2023.101283.
- [10] N. G. Goswami et al., "Detection of sickle cell disease using deep neural networks and explainable artificial intelligence," *J. Intell. Syst.*, 2024, doi: 10.1515/jisys-2023-0179.
- [11] B. S. D. Darshan et al., "Differential diagnosis of iron deficiency anemia from aplastic anemia using machine learning and explainable Artificial Intelligence utilizing blood attributes," *Sci. Rep.*, vol. 14, Art. 24120, 2024, doi: 10.1038/s41598-024-84120-w.
- [12] K. T. Navya et al., "Efficient diagnostic model for iron deficiency anaemia detection: a comparison of CNN and object detection algorithms in peripheral blood smear images," *Automatika*, vol. 66, no. 1, pp. 1–15, 2025, doi: 10.1080/00051144.2024.2433868.
- [13] M. Shahzad et al., "Identification of Anemia and Its Severity Level in a Peripheral Blood Smear Using 3-Tier Deep Neural Network," *Appl. Sci.*, vol. 12, no. 10, Art. 5030, May 2022, doi: 10.3390/app12105030.
- [14] R. Mantri, R. A. H. Khan, and S. Jadhav, "Leukemia diagnosis using transfer learning: An efficient approach," *Front. Health Inform.*, vol. 13, no. 2, 2024.
- [15] S. Dhengre et al., "Machine learning driven anemia identification and classification: A comprehensive survey," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 11, no. 11, Nov. 2023.
- [16] V. Gupta et al., "Relevance of peripheral blood smear examination in diagnosis of anemia in an era of automation," *Int. J. Acad. Med. Pharm.*, vol. 5, no. 4, Jul. 2023, doi: 10.47009/jamp.2023.5.4.198.
- [17] H. B. Baydargil and T. Bocklitz, "Unstained Blood Smear Analysis: A Review of Rule-Based, Machine Learning, and Deep Learning Techniques," *Journal of Biophotonics*, 2025, doi: 10.1002/jbio.202500057.
- [18] K. M. Aung, N. Z. Abdullah, N. A. Talib, M. Z. C. Azemin, and I. A. Bin Taib, "Detecting Red Blood Cell Morphology Changes In Iron Deficiency By Deep Learning Artificial Intelligence," *Revelation and Science: Special Issue Postgraduate Colloquium 2024 (1446H/2025)*, pp. 50-64.
- [19] D. Muduli et al., "Deep learning-based detection and classification of acute lymphoblastic leukemia with explainable AI techniques," *Array*, vol. 26, 2025, Art. no. 100397.
- [20] S. Kumar et al., "Enhanced Multi-Scale Convolutional Neural Network with Attention Mechanism for Accurate and Efficient Automated Hematological Diagnostics from Blood Smear," in *2025 International Conference on Electronics and Renewable Systems (ICEARS)*, Tuticorin, India, 2025, pp. 1750-1758, doi: 10.1109/ICEARS64219.2025.10940568.
- [21] O. El Othmani et al., "AI-driven Automated Blood Cell Anomaly Detection: Enhancing Diagnostics and Telehealth in Hematology," *J. Imaging*, vol. 11, no. 5, p. 157, 2025, doi: 10.3390/jimaging11050157.
- [22] D. A. Mura et al., "Exploring Few-Shot Object Detection on Blood Smear Images: A Case Study of Leukocytes and Schistocytes," *arXiv preprint arXiv:2503.17107*, 2025.

- [23] M. T. Mutar et al., "Optimizing deep learning for accurate blood cell classification: A study on stain normalisation and fine-tuning techniques," *Iraqi J. Hematology*, vol. 14, no. 1, pp. 60-65, 2025, doi: 10.4103/ijh.ijh_110_24.
- [24] S. Sahay, "Generalizable Blood Cell Detection via Unified Dataset and Faster R-CNN," arXiv preprint arXiv:2511.08465, 2025.
- [25] Y. Feng et al., "A Multi-Strategy Active Learning Framework for Enhanced Peripheral Blood Cell Image Detection," *IEEE Access*, vol. 13, pp. 104815-104827, 2025, doi: 10.1109/ACCESS.2025.3579918.
- [26] P. Appiahene, J. W. Asare, and E. T. Donkoh, "Detection of Iron Deficiency Anemia by Medical Images: A Comparative Study of Machine Learning Algorithms," *BioData Mining*, vol. 16, Art. 2, 2023, doi: 10.1186/s13040-023-00319-z.
- [27] I. N. Margret and K. Rajakumar, "Adaptive Elastic GAN for High-Fidelity Blood Cell Image Hallucination and Classification," *IEEE Access*, vol. 13, pp. 84897-84910, 2025, doi: 10.1109/ACCESS.2025.3568539.
- [28] O. El Othmani, S. Naouali, and H. S. Alkahtani, "Transformer Networks for Morphological Analysis and Functional Status Classification of Blood Cells," in *2025 IEEE 4th International Conference on Computing and Machine Intelligence (ICMI)*, MI, USA, 2025, pp. 1-6, doi: 10.1109/ICMI65310.2025.11141145.
- [29] A. Kumar and L. Nelson, "Deep Learning-based Blood Cell Classification using EfficientNetB3 Architecture," in *2025 6th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI)*, Goathgaun, Nepal, 2025, pp. 954-960, doi: 10.1109/ICMCSI64620.2025.10883188.
- [30] K. Naruenatthanaset, T. H. Chalidabhongse, D. Palasuwan, N. Anantrasirichai, and A. Palasuwan, "Red Blood Cell Segmentation with Overlapping Cell Separation and Classification on Imbalanced Dataset," arXiv preprint arXiv:2012.01321, 2020.
- [31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.
- [32] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Long Beach, CA, USA, 2019, pp. 6105-6114.
- [33] C. Szegedy et al., "Going Deeper with Convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, 2015, pp. 1-9, doi: 10.1109/CVPR.2015.7298594.
- [34] R. R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, 2017, pp. 618-626, doi: 10.1109/ICCV.2017.74.
- [35] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Long Beach, CA, USA, 2017, pp. 4765-4774.
- [36] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, San Francisco, CA, USA, 2016, pp. 1135-1144, doi: 10.1145/2939672.2939778.
- [37] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," arXiv preprint arXiv:2010.11929, 2020.
- [38] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Montreal, Canada, 2021, pp. 10012-10022, doi: 10.1109/ICCV48922.2021.00986.
- [39] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 779-788, doi: 10.1109/CVPR.2016.91.
- [40] T.-Y. Lin et al., "Feature Pyramid Networks for Object Detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, 2017, pp. 2117-2125, doi: 10.1109/CVPR.2017.106.
- [41] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, 2017, pp. 2961-2969, doi: 10.1109/ICCV.2017.322.

- [42] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in Proc. Med. Image Comput. Comput.-Assist. Interv. (MICCAI), Munich, Germany, 2015, pp. 234-241, doi: 10.1007/978-3-319-24574-4_28.
- [43] V. Macenko et al., "A Method for Normalizing Histology Slides for Quantitative Analysis," in Proc. IEEE Int. Symp. Biomed. Imaging (ISBI), Boston, MA, USA, 2009, pp. 1107-1110, doi: 10.1109/ISBI.2009.5193250.
- [44] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, "Color Transfer Between Images," IEEE Comput. Graph. Appl., vol. 21, no. 5, pp. 34-41, Sep./Oct. 2001, doi: 10.1109/38.946629.
- [45] Y. Ganin and V. Lempitsky, "Unsupervised Domain Adaptation by Backpropagation," in Proc. Int. Conf. Mach. Learn. (ICML), Lille, France, 2015, pp. 1180-1189.
- [46] I. Goodfellow et al., "Generative Adversarial Nets," in Proc. Adv. Neural Inf. Process. Syst. (NIPS), Montreal, Canada, 2014, pp. 2672-2680.
- [47] B. McMahan et al., "Communication-Efficient Learning of Deep Networks from Decentralized Data," in Proc. Int. Conf. Artif. Intell. Statist. (AISTATS), Fort Lauderdale, FL, USA, 2017, pp. 1273-1282.
- [48] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in Proc. Adv. Neural Inf. Process. Syst. (NIPS), Lake Tahoe, NV, USA, 2012, pp. 1097-1105.
- [49] M. J. Page et al., "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," BMJ, vol. 372, p. n71, 2021, doi: 10.1136/bmj.n71.
- [50] A. Vaswani et al., "Attention Is All You Need," in Proc. Adv. Neural Inf. Process. Syst. (NIPS), Long Beach, CA, USA, 2017, pp. 5998-6008.
- [51] N. Carion et al., "End-to-End Object Detection with Transformers," in Proc. Eur. Conf. Comput. Vis. (ECCV), Glasgow, UK, 2020, pp. 213-229, doi: 10.1007/978-3-030-58452-8_13.