Image and Vision Computing 162 (2025) 105673

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Distributed collaborative machine learning in real-world application scenario: A white blood cell subtypes classification case study

Lorenzo Putzu [a,*], Simone Porcu [a,b], Andrea Loddo [c]

[a] Department of Electrical and Electronic Engineering, University of Cagliari, Italy
[b] Consorzio Nazionale Interuniversitario per le Telecomunicazioni, University of Cagliari, Italy
[c] Department of Mathematics and Computer Science, University of Cagliari, Italy

ARTICLE INFO

Keywords:
Leukocyte classification
Domain shift
Domain generalisation
Federated learning
Early fusion
Late fusion

ABSTRACT

White blood cell (WBC) subtype classification is a critical step in monitoring an individual's health. However, it remains a challenging task due to the significant morphological variability of WBCs and the domain shift introduced by differing acquisition protocols across hospitals. Numerous approaches have been proposed to mitigate domain shift, including supervised and unsupervised domain adaptation, as well as domain generalisation. These methods, however, require a suitable amount of representative target images, even if unlabelled, or a suitable amount of images from multiple sources, which may not be feasible due to privacy regulations. In this study, we explore an alternative paradigm, known as *Distributed Collaborative Machine Learning* (DCML), which consists of exploiting images from different sources in a privacy-preserving setup. Although DCML methods seem well suited to this application, to the best of our knowledge, they have not been used for this task or to address the above-mentioned issues. However, we argue that DCML deserves further consideration in medical images as a potential alternative solution against domain shift in a privacy-preserving setup. To substantiate our view, we consider three DCML methods: early and late fusion and federated learning approaches, each offering distinct trade-offs in terms of training constraints, computational overhead and communications costs. We then conduct an extensive, cross-dataset experimental evaluation on four benchmark datasets and provide evidence that even *simple* implementations of DCML methods can effectively mitigate domain shift in WBC classification tasks.

1. Introduction

White blood cells (WBCs), medically defined as leukocytes, represent integral constituents within the immunological milieu, primarily protecting the host organism against microbial intrusion and pathogens. The classification of WBCs is essential in healthcare as it provides vital information for diagnosing diseases and guiding treatment decisions [1–3]. For example, we find a crucial example of how important WBC recognition is in the treatment of leukaemia, a group of blood cancers characterised by the uncontrolled proliferation of abnormal WBCs [4–6]. According to Siegel et al. leukaemia is estimated to cause around 23,710 deaths in the United States in 2023, emphasising the urgent need for efficient WBC identification technology [7].

WBCs are comprehensively categorised into five main subtypes: neutrophils, eosinophils, basophils, lymphocytes, and monocytes, each with specific functions in the immune system. Maintaining a standard WBC count is essential for overall well-being and health [2,3]. The taxonomy and classification of WBCs serve as essential milestones in the diagnostic and healing landscape of various haematologic conditions, encompassing but not limited to leukaemia, lymphoma, and infectious diseases. Conventional approaches to WBC subtypes classification, such as manual microscopy, exhibit proclivities towards protracted processing times and the concomitant susceptibility to human errors. Recent years have witnessed the rise of Machine Learning (ML) and, in particular, Deep Learning (DL) as a promising and transformative paradigm for WBC categorisation [8].

Based on extensive training datasets comprising meticulously labelled WBC images, DL algorithms acquire the ability to discern and categorise distinct leukocyte varieties autonomously [8,9]. This paradigm offers the potential to enhance the precision of WBC subtypes classification while mitigating the risk associated with human fallibility. However, automation of this procedure faces challenges, starting with clinical factors such as the diverse shapes and structures of cells and variations in protocols across different hospitals, including differences in equipment, lighting and staining procedures [10,11]. These

* Corresponding author.
E-mail addresses: lorenzo.putzu@unica.it (L. Putzu), simone.porcu@unica.it (S. Porcu), andrea.loddo@unica.it (A. Loddo).

https://doi.org/10.1016/j.imavis.2025.105673
Received 29 January 2025; Received in revised form 27 May 2025; Accepted 15 July 2025
Available online 11 August 2025

variations can result in highly distinct images, leading to a domain shift that significantly impacts the performance of DL models trained solely on data from a single site. This, in turn, renders established approaches ineffective at specific sites [11–13]. To address the domain shift issue, existing solutions are based on domain adaptation (DA) [14,15] or domain generalisation (DG) [16] methods, aimed at improving the robustness of the model in the target domain by exploiting data (even unlabelled) from the target domain, or extracting domain invariant features from more domains.

In a real-world application scenario, collecting and aggregating data from the target domain or more domains simultaneously is not feasible [10,17]. In fact, beyond the clinical field, it is crucial to highlight that the classification of WBC transcends the domain of privacy considerations. WBCs contain information of a personal nature, encapsulating details regarding the patient's genetic constitution and immunological status. Privacy concerns directly affect the creation of centralised and comprehensive datasets. Federated Learning (FL) is a compelling solution from the *Distributed Collaborative Machine Learning* (DCML) field as a direct consequence of privacy issues [18–20]. Traditional ML approaches often require centralising large datasets, posing individual privacy risks. In contrast, FL allows medical institutions and researchers to collaborate on WBC classification models without sharing raw patient data. Enabling the collaborative development of classification models across decentralised data sources fosters a new frontier in healthcare research where privacy and innovation can coexist [19–21]. However, FL has high costs, as it needs the training process to run simultaneously on multiple client nodes (those hosting the different datasets) and a server node that is in charge of aggregating/fusing the data [21,22]. In addition, during the entire training procedure, it is imperative to ensure the safeguarding of all nodes, fostering secure communication among them. Indeed, it has recently been highlighted that the data used during the learning process can be partially reconstructed from network weights or gradients, making this process particularly vulnerable to possible malicious attacks [23,24].

In the context of DCML, early fusion and late fusion are two strategies for enabling information exchange between neural networks without requiring simultaneous joint training. In early fusion, raw data from different modalities are combined prior to feature extraction, allowing the model to learn cross-modal correlations from the outset. However, this approach can be complex due to the challenge of modelling inter-modal relationships at such an early stage [25]. In contrast, late fusion involves independently extracting features from each modality, which are then combined, typically at the decision level, for classification. This results in a more modular and often more interpretable multi-modal representation [25].

This work aims to present the DCML paradigm as a potential alternative solution against domain shift in a privacy-preserving setup. We argue that DCML can be viewed as a kind of domain generalisation in different medical tasks and highlight the case of DCML for WBC subtypes classification. While early fusion, late fusion, and FL are well-established collaborative learning paradigms, to the best of our knowledge, no prior work has performed a direct empirical comparison of these approaches under a consistent and reproducible cross-dataset protocol tailored to real-world biomedical imaging constraints. In this study, we aim to fill this gap by systematically evaluating these three paradigms in the context of WBC subtypes classification using heterogeneous, decentralised data sources. Specifically, we design an experimental setup in which each dataset simulates an independent medical institution, and cross-dataset generalisation is assessed through a consistent source-target rotation protocol.

Our main contributions can be summarised as follows:

- We present an extensive comparative analysis of various approaches for WBC subtypes classification to enhance the understanding of the potential benefits and trade-offs associated with adopting the DCML approach in this domain. This comparison encompasses three DCML approaches: FL, early fusion, and late fusion. Additionally, the comparison includes the traditional centralised approach as a reference point since it is commonly employed in WBC subtypes classification tasks. To ensure comparable performance under each test condition, the same types of Convolutional Neural Network (CNN) and Vision Transformer (ViT) architectures are utilised to evaluate the four approaches.
- Our investigation explored the DCML paradigm, originally developed for tasks unrelated to WBC subtypes classification. We demonstrate its potential to effectively address domain shift challenges commonly encountered in WBC subtypes classification, such as staining variability, colour differences, and heterogeneous acquisition conditions, highlighting how DCML can mitigate biases introduced by these factors. These findings complement existing approaches in domain adaptation and generalisation, while simultaneously preserving patient privacy.
- We established a concrete baseline across four public WBC datasets through cross-dataset evaluations. To the best of our knowledge, this study represents the first comprehensive assessment of personalised and collaborative learning approaches applied to diverse WBC datasets. In particular, we conducted an in-depth analysis of the limitations of widely used DL models for WBC subtypes classification when exposed to data distributions that differ from those seen during training, highlighting the persistent challenge posed by domain shift. Moreover, our findings demonstrate the effectiveness of the FL paradigm in mitigating domain shift.
- We offer several visual insights with established techniques to show how the proposed approach succeeds in focusing on the discriminative zones of WBCs despite domain shifting, making it extremely competitive in this domain and task.

The remaining sections of the manuscript are structured as follows: Section 2 outlines relevant previous research on WBC analysis, particularly focusing on domain shift solutions and privacy-preserving collaborative methods. The latter are deepened in Section 3 while presenting the considered reference scenario. Section 4 describes the benchmark datasets, the DL architectures and the experimental setting used in our evaluation. The experimental results, a comparison with existing literature, and related discussions are presented in Section 5. Lastly, Section 6 provides conclusions and discusses potential avenues for future research.

2. Related work

This section provides an overview of the existing literature on WBC analysis, focusing on computer-aided systems (see Section 2.1) and domain shift solutions (see Section 2.2), as well as privacy-preserving methods (see Section 2.3).

2.1. WBC analysis using computer-aided systems

The diagnosis of haematopoietic malignancies, such as leukaemia, presents great hurdles due to the limited treatment options available. Traditionally, accurate diagnosis relies on the meticulous cytological evaluation of WBCs in blood or bone marrow smears. However, the advent of Computer-Aided Diagnosis (CAD) systems has ushered in a new era, promising to revolutionise this process by automating and enhancing diagnostic accuracy and efficiency [11,26–29].

Existing CAD systems for WBC analysis encompass a spectrum of tasks, ranging from simple cell counting to the detection and classification of diseases. Notably, these systems excel in classifying WBC subtypes [9,28,30–33] and identifying specific leukaemia conditions, such as acute lymphoblastic leukaemia (ALL) or acute myeloid leukaemia (AML) [8,11,34].

For instance, Matek et al. employed the ResNext architecture to accurately identify WBCs in AML patients' blood smears [28]. Acevedo

et al. proposed a predictive model for identifying patients with myelodysplastic syndrome, a precursor to AML [26]. Additionally, Vogado et al. leveraged transfer learning across multiple CNN architectures, combined with Support Vector Machine (SVM) classifiers, achieving noteworthy results in leukaemia diagnosis [29].

Furthermore, researchers have explored hybrid approaches, combining image processing and DL methodologies to achieve outstanding classification accuracy. For instance, Şengür et al. demonstrated the efficacy of such an approach in Whole Blood Cell Count (WBCC), underscoring the synergy between traditional and modern techniques [35], while Huang et al. developed a WBC classification framework by combining modulated Gabor wavelet kernels and deep CNN kernels, enabling the network to learn representative features at different frequencies and orientations [27].

As a general rule, the main advancements on the field are brought by the creation of novel architectures, either CNNs [31,36–39] or ViTs [9,40,41], ensemble strategies [42], or hybrid methods [32,33,43,44] potentially combining feature selection techniques [45].

A significant advancement in the field is exemplified by the development of BloodCaps by Long et al. a capsule-based model designed specifically for the accurate classification of various blood cell types in peripheral blood images. BloodCaps surpassed the performance of traditional CNNs, such as AlexNet, VGG-16, ResNet-18, and InceptionV3, underscoring the potential of novel architectural paradigms to improve classification techniques [37].

Other authors focused not only on performance but also on computational costs. For such a purpose, Firat et al. introduced a novel multi-branch lightweight CNN architecture featuring three parallel branches: the Inception module, depthwise squeeze-and-excitation block (DSEB), and pyramid pooling module (PPM). The Inception module enhances both efficiency and classification accuracy by conducting simultaneous convolutions at multiple scales. The DSEB selectively identifies informative features while eliminating redundant ones, thereby requiring minimal computational resources. Additionally, the PPM captures multi-scale contextual information from input images through feature pooling at various scales [31]. Methods aimed at training with limited data, such as few-shot learning approaches [38], have received limited attention due to the prevalence of large datasets in this domain, akin to the ones utilised in this study.

Regardless of the type of supervision used, novel architectural approaches in WBC classification might encounter common challenges inherent to this context, such as the dependency on specific datasets. This reliance can lead to substantial performance degradation when models are exposed to domain shifts—a central issue addressed in this study. Loddo et al. analysed the robustness of current WBCs classification methods in real application scenarios, characterised by the presence of imprecise regions of interest [11]. Their findings highlighted that even simpler approaches could yield accurate results when trained on curated benchmark datasets, but only a few approaches, in particular modern CNNs, exhibited robustness in handling real-world complexities.

Some of these challenges are intrinsic to cytopathological images that exhibit cells with diverse shapes and structures or disease-induced morphological alterations. Other challenges, instead, are introduced during the image acquisition process, where variability in equipment can introduce issues such as inconsistent lighting or blurred boundaries. Moreover, during smear preparation, variations in stain colour or staining protocols across different hospitals can further exacerbate the problem, contributing to substantial visual discrepancies between images from different sources.

The resulting domain shift can severely affect the generalisability of models trained on data from a single institution, often requiring re-annotation and model re-training for deployment in new environments. Addressing domain shift during model optimisation is thus crucial for ensuring cross-site applicability. Strategies such as DA [14,15] and DG [16] offer promising avenues for aligning disparate data domains in practical applications.

2.2. Overcoming domain shift challenges

DA is a key technique for cross-dataset evaluation, consisting of training a classifier on labelled data from a source domain and adapting it to a related, albeit shifted, target domain. In medical image analysis, many methods focus on aligning the feature distributions of the source and target domains to minimise the domain shift [14,16,46] using *target reconstruction* [47], *adversarial learning* [48,49], or *divergence minimisation* [50]. *Domain randomisation* is also widely used. It is based on the assumption that perturbing source domain images with variations such as colour changes, texture shifts, and noise addition helps reduce domain shift [51].

DG is a paradigm aiming to train models that generalise to unseen test domains. This is a challenging task, as it requires the model to learn to extract domain-invariant features that are representative of the underlying task, regardless of the specific domain [16,46]. There are two main approaches to DG. The first, *domain-invariant representation learning*, aims to learn a representation of the data that is invariant to different domains. This is typically done by minimising the domain discrepancy between the source domains, for example, with adversarial learning [52]. The second, *episodic learning*, aims to train the model to generalise to new domains by simulating virtual meta-tasks episodically. The model is trained on a sequence of episodes, each containing a small number of samples from different domains [53].

A few studies have addressed domain shift in WBC classification for multiple purposes, such as improving the accuracy of WBC classification on data from different domains or adapting models to new domains without forgetting previously learned knowledge [14,15,54,55]. Broadly, these approaches can be divided into two categories: (1) those that use knowledge of the target domain, and (2) those that use knowledge of all domains, including both source and target domains.

In the first category, Pandey et al. proposed a target-independent DA technique using CNNs to reduce domain shift between different camera types, a common challenge in WBC classification. Their approach comprises two stages: feature extraction, where a pre-trained CNN extracts source domain features, and domain adaptation, where a domain classifier aligns source and target domain feature distributions using domain-adversarial loss [14].

Umer et al. addressed the challenges of data imbalance, missing classes, and domain shifts that arise when training on datasets from multiple hospitals with different class distributions. Their proposed method uses a label-distribution-aware margin loss to handle imbalanced datasets and a domain-adaptive regularisation method to address domain shifts [55].

In the second category, Salehi et al. employed unsupervised cross-domain feature extraction for single blood cell image classification. Their method learned domain-invariant features from unlabelled images of the source domain using a deep convolutional autoencoder, and then used these features to classify images of the target domain [15].

Sadafi et al. proposed a continual learning approach for class incremental and domain incremental scenarios in WBC classification. Their method, based on a CNN, is trained on a source domain and fine-tuned on a target domain, ensuring adaptability to new domains and classes. It involves selecting the most confident samples and the most challenging samples identified through uncertainty estimation of the model [54].

Although DA and DG solutions are promising and have already yielded some level of results, it is necessary to consider that (1) access to the target domain or knowledge of all domains involved is typically not a real-world application case and (2) sharing medical imaging data across multiple centres is challenging due to privacy protection policies, such as Health Insurance Portability and Accountability Act (HIPAA) [56] and General Data Protection Regulation (GDPR) [57]. Therefore, new approaches are needed to create accurate models without accessing private patient information or transmitting raw data.

2.3. Privacy-preserving methods

Due to privacy considerations related to medical data, directly sharing, pooling or merging medical images across different medical centres is typically not feasible in real-world practice [58,59]. Consequently, privacy-preserving solutions are essential for medical research, allowing researchers to collaborate on sensitive data without sharing or leaking it.

Among privacy-preserving methods, FL excels as the latest DCML technique that allows multiple parties to train a shared model collaboratively without disclosing their raw data. In detail, the training process involves n clients, each possessing a comparable type of labelled data. These clients actively participate in training a unified predictive model. Employing an identical neural network architecture, each client conducts a single training epoch using its specific dataset. A central server, responsible for coordinating the training process, gathers and consolidates all the weights from the individual clients. Subsequently, the server disseminates the updated weights back to the clients, who use them to resume the training. This process is repeated until convergence is attained [60–62].

FL has been shown to be a promising approach for DCML in the medical field, as it allows researchers to train models on sensitive patient data without compromising privacy. FL has also been shown to perform comparably to centralised training on medical datasets, which means that it can be used to train accurate and robust models without the need to share the raw data [58,59].

In the realm of WBC classification, only two studies have utilised FL methodologies [63,64]. Madni et al. proposed a framework that can tackle both model and data heterogeneity by incorporating knowledge distillation and symmetric loss, respectively. The authors conducted experiments utilising the INT_20 dataset [65] and the AML dataset [28] (see Section 4.2 for details). The INT_20 dataset served as a publicly available dataset hosted on a server, while the AML dataset was distributed to individual clients as their local private datasets. Each experiment involved the configuration of four clients for collaborative learning, with the AML dataset being evenly partitioned among these clients using a Dirichlet distribution to create non-independent and identically distributed (non-IID) data [63].

A similar experiment was conducted by Zhu et al. but on the PBC dataset (see Section 4.2 for details), wherein the non-IID data are distributed among eight clients. They configured the server with a test set comprising both known and unknown classes to evaluate the performance of the global model. This global model was formed by aggregating the outputs of multiple clients trained on known classes. Their objective was to accurately classify known classes and identify unknown classes as a separate class, whereas our study focuses on recognising five distinct WBC subtypes across different datasets, accounting for domain shifts.

Regardless of the final classification task, these two papers clearly focused only on simulations where a single dataset was distributed among multiple clients, thereby overlooking the privacy concern entirely. Furthermore, their proposed method relies on a feature extractor that is fine-tuned specifically for a single dataset, which limits its generalisability [64]. Therefore, further analysis is still needed to determine the generalisation ability of DCML methods to address the WBC classification task.

3. Revisiting DCML for domain generalisation

This section aims to provide a comprehensive understanding of the considered scenario and the DCML paradigm. Additionally, we argue for the applicability of DCML in WBC subtypes classification, particularly in scenarios where privacy constraints are stringent.

3.1. Reference scenario

As the existing literature indicates (see Section 2.1), prevailing methods for WBC subtypes classification typically rely on traditional centralised training with fully supervised setups using benchmark datasets. In this conventional approach, neural networks are trained on a singular set of fully labelled images and subsequently evaluated on a distinct partition of the same dataset, a setting known as the same-dataset evaluation setting. This method simulates a scenario wherein images are collected and labelled within a single haematological research centre. While this approach may yield exemplary performance on benchmark datasets [11], it remains somewhat detached from reality. Indeed, in practical application scenarios, acquiring labelled WBC images from the *target* haematological research centres—those that will employ the WBC subtypes classification system—is often unfeasible. Consequently, in this case, the model obtained from the traditional learning process must be utilised in a cross-dataset setting. However, benchmark datasets, despite their numerous variations, are subject to significant dataset biases, similarly to other computer vision tasks [66], thereby significantly impacting cross-dataset performance [11].

To address this challenge, researchers have proposed *unsupervised domain adaptation* (UDA) methods [15,48], wherein a model is trained on labelled data from a source domain (e.g., a benchmark dataset) and unlabelled data from a target domain (e.g., the target haematological research centre). However, UDA necessitates a substantial quantity of representative target images, a requirement that may be untenable for the considered task due to privacy protection policies. In contrast, DG eschews the utilisation of target data, instead aiming to enhance the model's generalisation capacity across any target domain. Generally, DG techniques learn domain-invariant features representative of the underlying task by leveraging multiple source domains for training without modifying the model post-deployment [16,46]. Nonetheless, collecting multiple source datasets in a centralised server may prove infeasible due to privacy protection policies.

The DCML paradigm emerges as an attractive solution in the aforementioned application scenario. Its inherent emphasis on utilising multiple source domains for training underscores the collaborative efforts of various haematological research centres within a privacy-preserving framework. Furthermore, we posit that this collaborative approach facilitates the extraction of domain-invariant features that accurately represent all involved domains, thereby promoting generalisation akin to DG. However, we emphasise that DCML represents a *complementary* approach to other DG techniques or methods against domain shift. Additionally, DCML can be seamlessly integrated with any DL architecture, rendering it model-agnostic.

3.2. DCML paradigm

In contrast to the traditional centralised approach, DCML allows for the collaborative training of diverse models without centralising the datasets on a single server. This framework enables the implementation of robust privacy policies and fosters collaboration among entities responsible for managing private data, thereby facilitating the development of high-performance predictive models. Existing DCML methodologies exhibit variability not only in performance but also in associated costs pertaining to training, inference, communications, and security. In this study, we conduct a comprehensive evaluation, including a critical analysis of cost implications, by considering three DCML techniques: **early fusion**, **late fusion**, and **Federated Learning**. It must be noted that we considered only DCML techniques designed to enhance the robustness of ML models by aggregating various information sources in a privacy-preserving environment. In this context, we did not explore techniques such as Split Learning, which, although operating within a privacy-preserving framework, aims to minimise client-side computations of low-end edge devices by directing them to a powerful central server [67].

Early fusion and late fusion techniques originate from the multimodal domain, wherein various entities collaborate to predict a common target [25,68–70]. The modalities correspond to different haematological research centres in this context, each training their DL model on their respective datasets using a traditional, fully supervised setup. In a subsequent phase, wherein model training need not occur simultaneously, each research centre shares its model weights with a dedicated server, responsible for averaging and redistribution operations. In detail, in the case of early fusion, the server averages the weights of all layers, resulting in a single consolidated model that is redistributed to all the research centres involved. The late fusion approach instead entails the continued use of original models since fusion occurs at the decision-making level. The server only takes care of creating an additional layer that can mediate the decisions generated by the n models (considering n entities), yielding a new model with multiple branches, and finally redistributing the other $n-1$ branches to each entity. The late fusion aims to capitalise on the complementary information gleaned from individual models to be fully leveraged during inference. However, this collaborative approach incurs higher costs associated with the number of participating entities, which determines the final number of branches. It is important to highlight that early fusion presupposes the use of homogeneous DL architectures across all entities, whereas late fusion offers greater flexibility, requiring a shared structure only for the last layer [68]. For the sake of comparability, we exclusively employ homogeneous architectures in this context.

It must also be noted that, in early and late fusion, each client trains its model independently and communicates asynchronously with a central server after the training process. No synchronisation is required between clients, which makes these methods communication-efficient but less interactive. In contrast to early and late fusion techniques, FL operates on a completely decentralised learning paradigm using synchronous communication. It enables multiple parties to train a single DL model collaboratively without directly sharing raw data. The core principle involves training n local models in parallel, each with its private dataset, and sharing the models' parameters (i.e., weights) with a central server after each round. The server is responsible for averaging the received weights and synchronously redistributing the updated model to all clients, requiring coordination at each communication round [20,60–62]. Naturally, this approach incurs heightened communication and computation costs, necessitating the establishment of secure and prolonged communication channels between clients and the server. In particular, each training round entails multiple bidirectional transmissions of model parameters or gradients, and the need for synchronous updates may exacerbate delays or inefficiencies in the presence of unreliable or slow clients.

Beyond privacy and synchronisation overheads, communication efficiency and system scalability represent critical factors in FL deployments, especially in medical settings where bandwidth or compute power may vary across centres. While our work focuses on model performance and generalisation under distribution shift, we recognise the relevance of system-level trade-offs. To illustrate the cost of model transmission in practice, we report the total communication overhead observed in a recent analysis of different FL algorithms [71]. Table 1 shows the cumulative data exchanged over the network (in megabytes) to reach various accuracy thresholds, measured via the Matthews Correlation Coefficient (MCC), using 8 clients. Despite using full model weight updates, the total volume remained modest, highlighting the communication efficiency of this approach under limited rounds and shallow models.

These values reflect full-model updates per round, averaged across clients, and are consistent with scenarios where communication is performed only every few local epochs. In more complex or asynchronous settings, such as in edge clinical deployments, further reductions may require sparse updates or quantisation-aware optimisation. For a detailed quantitative analysis of communication rounds, bandwidth requirements, and optimisation techniques such as update compression or client selection, we refer the reader to [20–22,71]. These considerations are particularly pertinent in clinical scenarios involving large models or geographically distributed nodes, where reducing the communication footprint is often essential for practical deployment.

Table 1
Total communication overhead (in MB) for decentralised synchronous FL to reach target MCC thresholds.

Target MCC	5G Slicing (MB)	IoT-DNL (MB)
0.70	2.03	1.76
0.80	2.61	3.51
0.90	4.05	14.94
0.95	4.83	32.51

Additionally, FL encounters challenges stemming from heterogeneity in data distribution and system characteristics across participating parties, which further complicate coordination and efficiency. To mitigate these challenges, various federated optimisation methods, such as FedAVG and FedProx, have been proposed. FedAVG involves averaging local model updates directly at the central server, while FedProx introduces a proximal term to the local objective function, handling local data and computation variability. Both methods have demonstrated favourable performance and convergence properties across diverse FL scenarios [72,73]. The straightforward nature of FedAVG may lead to expedited convergence in scenarios where the added intricacy of FedProx is unnecessary, particularly in cases where the data is unaffected by non-IID issues. Additionally, it is worth noting that FedProx necessitates more sophisticated hardware due to the increased computational overhead on devices, thereby augmenting the inherent cost associated with implementing this DCML methodology [74].

4. Experimental evaluation

In this section, we present the DL architectures used in this study (see Section 4.1), explore benchmark datasets (detailed in Section 4.2), and outline the experimental setup (see Section 4.3).

4.1. Deep learning architectures

Our experimentation's primary objective revolves around assessing and juxtaposing various DL architectures for the classification of WBC subtypes, aiming to ensure the robustness of our findings against network-specific intricacies such as layer counts, skip connections, and other architectural nuances. While specialised architectures tailored explicitly for WBC classification exist [9,37], characterised by distinct supervisory mechanisms and domain-specific features, we opted for more generic and widely-adopted classification architectures to facilitate broader applicability and comparative analysis. To comprehensively cover the architectural landscape, we deliberately selected one representative model from each of the main DL architecture families: residual networks (ResNet-152, from now on Res.152), lightweight networks (MobileNetV3, from now on Mob.v3), densely connected networks (DenseNet-121, from now on Den.121), inception networks (InceptionV3, from now on Inc.v3), as well as transformer-based models, including the Vision Transformer (ViT) and and hierarchical vision transformers (Swin).

This selection enables us to investigate the effectiveness and generalisability of DCML paradigms across models with diverse inductive biases, computational complexities, and performance characteristics. At the same time, all of them have proven their ability to extract deep hierarchical features and are widely adopted in medical imaging benchmarks [9,31,36–41,75].

The details of the used architectures are as follows. **Res.152**, belonging to the ResNet family of architectures, comprises 152 layers

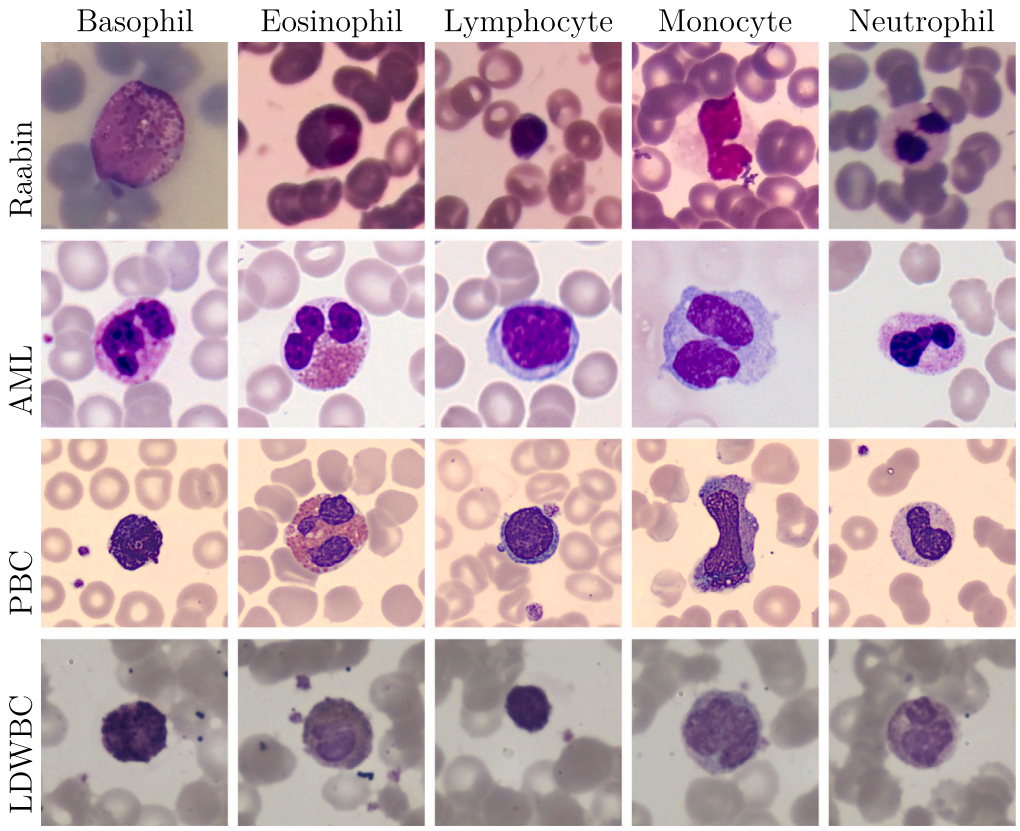

Fig. 1. Comprehensive overview of the four datasets utilised in this study: Raabin, AML, PBC, and LDWBC. The figure delineates the five adopted classes: basophil, eosinophil, lymphocyte, monocyte, and neutrophil.

and is emblematic of residual learning principles [76]. These architectures incorporate skip connections or recurrent units between convolution and pooling layer blocks, coupled with batch normalisation post blocks [77].

Mob.V3 is a small architecture recognised as the latest and most compact variant among the MobileNet family. This family features lightweight CNN architectures tailored for mobile and embedded devices. These models leverage depthwise separable convolutions to replace traditional convolution layers, thereby reducing computational overhead and model size without sacrificing accuracy [78].

Inc.v3, a successor to GoogLeNet, is founded on Inception layers, wherein each block constitutes a set of convolution layers accommodating filters ranging from $1x1$ to $5x5$. This design facilitates learning across multiple scales and adopts global average pooling over traditional max-pooling, alongside factorised, smaller, and asymmetric convolutions, supplemented by an auxiliary classifier for network regularisation [79,80].

Den.121 is part of the DenseNet family, introduced to address the constraint in traditional CNNs where the number of layers (L) equals the number of connections. This architecture introduces $L(L+1)/2$ connections, wherein the output of each layer serves as input to all subsequent layers. The number of filters in each convolution layer dynamically varies based on the growth rate parameter, k, thereby influencing the overall parameter count; k is set to 32 in the original work [81].

ViT stands as the pioneering transformer-based architecture proposed for CV tasks [82]. It leverages a self-attention mechanism, supplanting traditional convolution layers with self-attention layers. Images are partitioned into non-overlapping patches of fixed size and processed sequentially through transformer layers, thus furnishing spatial information to the model.

Swin, an advancement over the ViT model, addresses the challenge of long-range dependencies by employing a hierarchical self-attention mechanism. It adopts a shifted window approach for patch division, allowing overlapping patches to capture more contextual information. The resulting features are aggregated using Swin blocks, thereby encompassing both local and global dependencies [83].

4.2. Datasets

In our study, we deliberately opted to use four well-established and publicly available datasets to ensure the integrity and generalisability of our results across diverse data compositions. Several datasets for WBC subtype classification exist within the literature, yet certain limitations hinder their suitability for our comparative analysis. For instance, the LISC dataset [84], while commonly employed in related studies [12,13,85], exhibits a notably low image count for each WBC subtype, typically around 50 images per class. Similarly, the BCCD dataset lacks representation of basophiles, further limiting its applicability in our investigation [13].

To circumvent these constraints and foster a comprehensive evaluation framework, we selected datasets encompassing a broader spectrum of WBC classes with substantial image counts. Subsequently, we eliminated or grouped the classes in advance according to the classes pertinent to our study. Such a selection strategy aims to mitigate biases stemming from dataset composition and to facilitate a robust comparison among the performance of different classification approaches.

The **Raabin-WBC** (from now on Raabin) is an open-access dataset obtained from samples originating from different laboratories in Iran, with each sample undergoing Giemsa staining. The obtained images were subject to labelling and processing to create various subsets tailored for specific tasks. Our primary focus was on a subset that comprises 1,145 single-cell images standardised to dimensions of 575×575 pixels for the five main WBC classes: lymphocytes (242), monocytes (242), neutrophils (242), eosinophils (201), and basophils (218) [5].

The **AML-Cytomorphology-LMU** (from now on AML) dataset presents a comprehensive single-cell collection encompassing leukocytes from both AML patients (100) and non-malignant controls (100) collected at Munich University Hospital between 2014 and 2017. It comprises 18,365 expert-identified single-cell images with dimensions 400 × 400 categorised into 15 distinct single-cell image classes, with four dedicated to leukaemic cells and the remaining eleven representing healthy WBCs [28]. For this study, we focused on the five main WBC classes. To this end, we grouped 8,484 segmented neutrophils and 109 band neutrophils under a unified neutrophil class, and merged 3,937 typical lymphocytes with 11 atypical lymphocytes into a single lymphocyte class. The resulting class distribution is as follows: neutrophils (8,593), lymphocytes (3,948), monocytes (1,789), eosinophils (424), and basophils (79).

The **PBC** dataset comprises 17,092 single-cell images collected at the Core Laboratory of Hospital Clinic of Barcelona from 2015 to 2019, using the CellaVision DM96 analyser. The obtained images, sized 360 × 363 pixels, have been categorised meticulously annotated by expert clinical pathologists into eight groups representing normal WBCs: neutrophils, eosinophils, basophils, lymphocytes, monocytes, immature granulocytes, erythroblasts, and thrombocytes. For the purpose of this study, we considered again the five main WBC classes: neutrophils (3,329), eosinophils (3,117), basophils (1,218), lymphocytes (1,214), and monocytes (1,420) [30,86].

The **LDWBC** dataset consists of 22,645 single-cell images that originate from 150 microscope slides belonging to 150 healthy individuals. The blood samples were stained using Wright-Giemsa and automatically scanned by a light microscope equipped with OLYMPUS BX41 and OLYMPUS Plan *N* 100x/1.25 to generate the corresponding digital images [75]. Finally, the single WBCs were cropped into images sized 1280 × 1280 pixels and categorised as follows: basophil (224), monocyte (968), eosinophil (539), neutrophil (10,469), and lymphocyte (10,445).

A sample image for each dataset and each WBC subtype is shown in Fig. 1.

4.3. Experimental setup

The DL architectures utilised in this study were initially pre-trained on the widely recognised natural image dataset, ImageNet [87], before being adapted to medical image tasks. This adaptation process followed a well-established procedure for transfer learning and fine-tuning DL models, as outlined in prior literature [88]. During this adaptation, all DL layers were preserved except for the final fully-connected layer, which was replaced with a new layer initialised and configured to accommodate the specific object categories relevant to our study.

To simulate real-world scenarios, each of the aforementioned datasets was conceptualised as a distinct haematological research centre, housing patient-related data. The performance of traditional centralised, fully supervised setups was compared with that of DCML techniques by forming collaborative groups consisting of pairs of datasets. A visual representation of the experimental design is presented in Fig. 2. In this setup, models involved in early and late fusion were trained independently on local data and later transmitted to a central server using asynchronous communication for post-hoc aggregation—model weights in the case of early fusion, and predictions in the case of late fusion. More in detail, in the early fusion setup, each client trains an identical model architecture on its local dataset independently. After local training, the models' parameters (PyTorch state_dicts) are aggregated by computing the element-wise average of corresponding weights: $\theta_{\text{fused}} = \frac{1}{2}(\theta_A + \theta_B)$ In the late fusion setup, models are still trained independently on local data, but after local training, they are used to create a new model with multiple branches and with a new final layer that averages the probability distribution over the classes produced by each branch, using the following equation: $p_{\text{fused}}(x) = \frac{1}{2}\left(p_A(x) + p_B(x)\right)$ (x) are the softmax outputs of models A and B, respectively. In contrast, models in the FL setup were trained concurrently, following a synchronous update scheme using the FedAVG algorithm.

Cross-dataset experiments were conducted to assess the generalisation capabilities of the DCML approaches under consideration. Each dataset was alternately designated as the target (testing) dataset in these experiments, while the others were combined to form the source (training) set. To ensure a fair comparison and enable direct comparison on the same testing partition, all mentioned datasets were partitioned into three fixed subsets: training, validation, and testing sets, constituting 60%, 20%, and 20% of the original dataset sizes, respectively. A stratified sampling procedure was employed to maintain the original data distribution. Given the inherent class imbalance present in several of the datasets, such as the limited number of basophils in AML (only 79 instances), we applied a weighted random sampling strategy during training. This ensured that each mini-batch contained a balanced representation of all classes, regardless of their frequency in the original data. This reweighting was applied exclusively to the training set, so as not to influence validation or test metrics. Furthermore, to reduce overfitting risks and increase generalisability, we applied a data augmentation pipeline, also restricted to the training set, which included random rotation, horizontal and vertical flipping, centre cropping, Gaussian blurring, and colour jitter.

All experiments were conducted on a single machine featuring an Intel(R) Xeon(R) CPU E5-2670 v3 @2.30 GHz CPU, 128 GB RAM, and NVIDIA Quadro 24 GB GPU. Training processes were executed for a maximum of 100 epochs, with early stopping based on validation performance. A batch size of 32, Stochastic Gradient Descent (SGD) optimiser, and a learning rate of 0.001 were employed throughout the training procedures. Although the FL experiments were conducted on a single physical machine, they were implemented using the Flower framework [89], which provides a modular and extensible infrastructure for both simulated and real-world FL deployments. This choice enables a faithful orchestration of client–server interactions, including weight transmission, aggregation, and synchronous updates, thereby offering a realistic simulation of federated training dynamics. While the current setup does not model real-world issues such as variable network conditions, communication delays, or system unreliability, it offers a controlled environment for benchmarking collaborative learning strategies. The use of Flower ensures that our implementation remains directly transferable to a distributed deployment scenario.

5. Experimental results

In this section, we present the outcomes of our experimentation, divided into four subsections. We begin with a quantitative analysis in Section 5.1, where we provide numerical assessments of the performance metrics obtained from our models. Following this, in Section 5.2 we delve into a qualitative analysis, offering insights into the visual aspects and interpretability of our results. Next, we conduct a comprehensive comparison with the state of the art, juxtaposing our findings with existing methodologies in the field (see Section 5.3). Finally, we discuss the results in Section 5.4, wherein we critically analyse and interpret the implications of our results in the context of WBC classification.

5.1. Quantitative analysis

The results obtained on the benchmark datasets using the considered DL architectures are presented in Tables 2–5, where performance metrics including Accuracy (A), Precision (P), Recall (R), and F1-score (F1) are reported. These metrics are calculated individually for each class and subsequently aggregated using a weighted average procedure to derive a consolidated performance measure.

Remarkably, DCML approaches, with very few exceptions, exhibit superior performance compared to the centralised approach. Upon

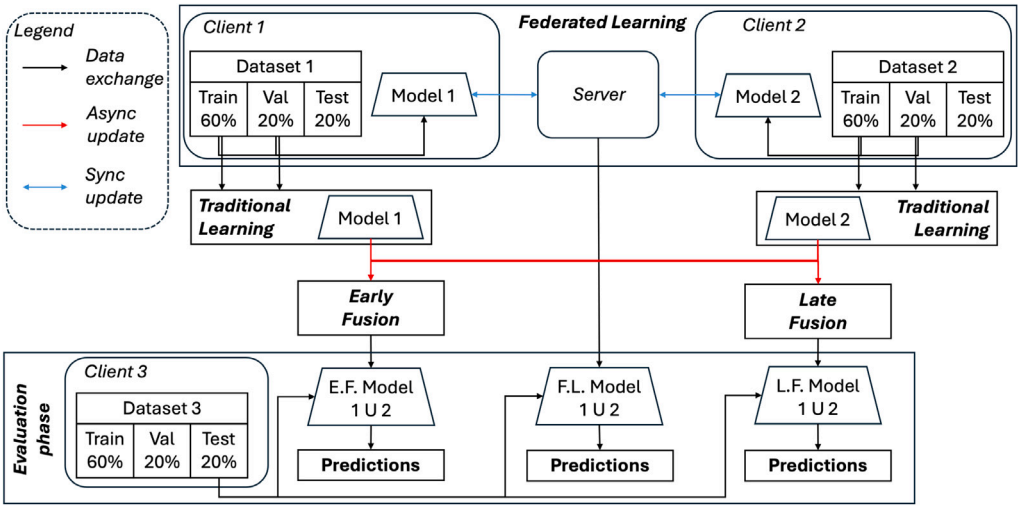

Fig. 2. Visual representation of the experimental design.

Table 2
Cross-dataset performances (Accuracy - A, Precision - P, Recall - R and F1-score - F1) on the Raabin dataset of the CL techniques compared to the centralised approach. The column SOURCE reports the dataset (or datasets, indicated as D1 ∪ D2) contributing to the final model. The best performance for each architecture is highlighted in boldface.

	Traditional learning					Collaborative learning													
	SOURCE	Centralised learning				SOURCE			Early fusion				Late fusion				Federated learning			
		A	P	R	F1				A	P	R	F1	A	P	R	F1	A	P	R	F1
Res.152	AML	64.2	79.9	64.2	56.4	AML	∪	PBC	65.5	**81.6**	65.5	58.7	65.9	79.7	65.9	59.0	**66.8**	81.0	**66.8**	**61.0**
	PBC	63.3	67.8	63.3	60.2	AML	∪	LDWBC	59.4	**82.4**	59.4	50.2	58.1	51.3	58.1	48.3	**72.5**	74.8	**72.5**	**70.7**
	LDWBC	44.5	43.5	44.5	36.0	PBC	∪	LDWBC	56.8	65.6	56.8	53.0	**59.8**	66.4	**59.8**	55.2	59.0	**72.4**	59.0	**60.9**
Mob.V3	AML	53.7	58.4	53.7	50.9	AML	∪	PBC	59.4	**68.5**	59.4	50.0	56.8	57.6	56.8	49.0	**59.8**	62.5	**59.8**	**51.3**
	PBC	47.6	38.9	47.6	38.7	AML	∪	LDWBC	**57.2**	58.0	**57.2**	53.8	**57.2**	**61.7**	**57.2**	**55.8**	44.5	45.8	44.5	44.0
	LDWBC	32.3	47.7	32.3	31.8	PBC	∪	LDWBC	45.9	47.8	45.9	42.3	48.5	45.4	48.5	44.0	**51.5**	**52.4**	**51.5**	**49.1**
Inc.v3	AML	62.9	73.7	62.9	59.3	AML	∪	PBC	57.2	71.5	57.2	49.9	65.1	**76.7**	**65.1**	**61.1**	**63.8**	70.7	63.8	55.3
	PBC	65.5	74.5	65.5	62.8	AML	∪	LDWBC	50.2	59.1	50.2	42.4	55.0	67.5	55.0	49.4	**73.8**	**79.9**	**73.8**	**69.9**
	LDWBC	45.0	51.0	45.0	37.6	PBC	∪	LDWBC	45.9	48.1	45.9	40.5	58.1	**70.3**	58.1	54.6	**68.1**	69.5	**68.1**	**65.9**
Den.121	AML	40.2	60.4	40.2	31.9	AML	∪	PBC	45.4	57.8	45.4	41.5	52.8	59.3	52.8	46.1	**64.6**	**82.9**	**64.6**	**58.1**
	PBC	63.8	67.6	63.8	60.5	AML	∪	LDWBC	35.8	40.4	35.8	29.4	46.7	60.9	46.7	37.7	**67.2**	**78.6**	**67.2**	**64.9**
	LDWBC	51.5	62.0	51.5	45.3	PBC	∪	LDWBC	49.3	59.7	49.3	46.5	**59.8**	**69.1**	**59.8**	**57.8**	56.3	62.6	56.3	54.3
ViT	AML	65.5	67.0	65.5	61.6	AML	∪	PBC	64.6	65.7	64.6	61.0	68.1	71.5	68.1	64.7	**72.9**	**79.7**	**72.9**	**68.7**
	PBC	68.1	**74.0**	68.1	63.9	AML	∪	LDWBC	66.4	68.2	66.4	64.1	66.8	70.1	66.8	63.9	**71.6**	70.9	**71.6**	**67.3**
	LDWBC	59.8	63.6	59.8	57.9	PBC	∪	LDWBC	**69.4**	**76.2**	**69.4**	**67.4**	68.1	72.9	68.1	64.5	73.4	76.4	73.4	**71.8**
Swin	AML	63.3	66.4	63.3	55.6	AML	∪	PBC	62.4	76.5	62.4	52.9	65.9	**79.3**	65.9	58.0	**72.9**	78.2	**72.9**	**70.1**
	PBC	67.7	77.1	67.7	61.5	AML	∪	LDWBC	68.6	79.6	68.6	62.2	69.4	**81.5**	69.4	63.9	**77.3**	79.7	**77.3**	**77.1**
	LDWBC	71.2	74.2	71.2	69.6	PBC	∪	LDWBC	**74.7**	**80.0**	**74.7**	**72.2**	72.5	79.5	72.5	69.0	70.7	77.0	70.7	70.1

comparing the three DCML approaches, FL consistently achieves superior performance. Moreover, FL demonstrates a smaller decline when the PBC dataset is the target. In this scenario, both early and late fusion approaches show a decrease in performance comparable to the centralised method, whereas FL displays a less pronounced decrease.

To further highlight these trends, Fig. 3 presents a box plot depicting the accuracy values obtained by all DL models trained using both centralised and DCML approaches, tested across the four target datasets. The efficacy of DCML methods in narrowing the confidence interval of predictions, thereby yielding more stable outcomes, is noteworthy. Notably, FL demonstrated superior performance across most instances, except for results obtained on the Raabin dataset, despite achieving the highest accuracy values.

A salient observation is the enhancement of model generalisation achieved through DCML methods, leading to heightened robustness in predictions. Notably, early fusion demonstrated advantages over the centralised approach; however, its static strategy for combining information limits its effectiveness in cross-dataset evaluation scenarios. In contrast, the most promising results were obtained using late fusion and FL approaches. The latter, in particular, benefits from iterative training across distributed clients, enabling the integration of diverse data representations over time, which consistently enhances classification performance.

The outcomes presented in Tables 2–5 unveil a notable trend: when incorporating the PBC dataset within the source, there is an observable enhancement in performance. This observation suggests that the PBC dataset encompasses considerable variability and possesses the capability to effectively represent diverse image acquisition systems/methodologies, thus exhibiting heightened potential for generalisation. Conversely, when the PBC serves as the target dataset, performance demonstrates increased variability, as exemplified more distinctly in Fig. 3.

Another consideration is that across most cross-dataset evaluations, FL consistently achieves the highest performance across all metrics, including recall, F1-score, and precision. Naturally, the outcomes are contingent upon the specific training datasets and the employed DL model, and variations arise when altering either of these parameters. Furthermore, under the test condition involving the InceptionV3 neural network trained with the Raabin and LDWBC datasets and tested on the PBC dataset, objective metrics indicate that one of the FL's relatively inferior results (A= 63.6%,P = 57.7%, R=63.6%, and F1=56.3%)

Table 3
Cross-dataset performances (Accuracy - A, Precision - P, Recall - R and F1-score - F1) on the AML dataset of the CL techniques compared to the centralised approach. The column SOURCE reports the dataset (or datasets, indicated as D1 ∪ D2) contributing to the final model. The best performance for each architecture is highlighted in boldface.

	Traditional learning					Collaborative learning													
	SOURCE	Centralised learning				SOURCE			Early fusion				Late fusion				Federated learning			
		A	P	R	F1				A	P	R	F1	A	P	R	F1	A	P	R	F1
Res.152	Raabin	76.8	77.7	76.8	77.1	Raabin	∪	PBC	71.1	79.7	71.1	70.3	**84.0**	**85.6**	**84.0**	**84.0**	78.4	85.1	78.4	78.5
	PBC	73.2	80.5	73.2	72.9	Raabin	∪	LDWBC	71.1	77.6	71.1	69.9	**76.3**	80.1	**76.3**	**74.8**	72.7	**81.2**	72.7	72.4
	LDWBC	66.5	78.7	66.5	62.8	PBC	∪	LDWBC	83.0	87.3	83.0	82.6	**86.1**	**87.5**	**86.1**	**85.7**	83.5	85.1	83.5	83.4
Mob.V3	Raabin	44.3	65.6	44.3	43.9	Raabin	∪	PBC	74.7	77.9	74.7	75.7	66.0	76.1	66.0	62.6	**82.5**	**83.7**	**82.5**	**82.6**
	PBC	**69.1**	**73.9**	**69.1**	**66.6**	Raabin	∪	LDWBC	56.7	62.5	56.7	54.1	54.1	53.0	54.1	50.7	64.9	73.3	64.9	64.3
	LDWBC	52.1	47.5	52.1	46.3	PBC	∪	LDWBC	**72.7**	**75.1**	**72.7**	**72.9**	63.4	52.1	63.4	56.5	70.1	75.0	70.1	68.5
Inc.v3	Raabin	71.1	77.7	71.1	73.3	Raabin	∪	PBC	83.0	84.2	83.0	83.4	**84.5**	**86.6**	**84.5**	**84.9**	77.8	84.2	77.8	77.7
	PBC	77.8	81.3	77.8	76.7	Raabin	∪	LDWBC	75.3	77.2	75.3	73.3	**80.4**	**82.0**	**80.4**	**79.8**	70.1	78.1	70.1	69.0
	LDWBC	67.5	75.6	67.5	65.5	PBC	∪	LDWBC	74.7	79.8	74.7	74.5	**79.4**	**83.5**	**79.4**	**78.5**	75.3	82.4	75.3	74.8
Den.121	Raabin	77.8	77.7	77.8	76.0	Raabin	∪	PBC	85.1	85.6	85.1	84.3	87.1	88.1	87.1	86.7	**90.7**	**90.9**	**90.7**	**90.7**
	PBC	74.7	80.7	74.7	74.4	Raabin	∪	LDWBC	76.8	81.1	76.8	75.9	**83.0**	**85.2**	**83.0**	80.9	82.5	84.4	82.5	**81.8**
	LDWBC	78.4	82.4	78.4	75.8	PBC	∪	LDWBC	81.4	85.7	81.4	81.2	**85.6**	**88.0**	**85.6**	**85.1**	85.1	87.2	85.1	**85.1**
ViT	Raabin	56.7	67.1	56.7	56.7	Raabin	∪	PBC	71.6	75.3	71.6	72.1	62.9	74.8	62.9	62.3	**83.5**	**85.5**	**83.5**	**83.9**
	PBC	64.9	**79.0**	64.9	65.2	Raabin	∪	LDWBC	66.0	74.3	66.0	64.4	63.4	74.3	63.4	61.7	**69.6**	76.1	**69.6**	**69.2**
	LDWBC	58.8	70.0	58.8	56.0	PBC	∪	LDWBC	68.6	76.4	68.6	67.5	64.9	75.7	64.9	64.1	**75.8**	**82.2**	**75.8**	**76.1**
Swin	Raabin	62.9	70.2	62.9	63.3	Raabin	∪	PBC	68.6	72.6	68.6	67.6	68.6	75.8	68.6	67.5	**82.0**	**84.6**	**82.0**	**82.2**
	PBC	69.6	75.8	69.6	67.5	Raabin	∪	LDWBC	60.8	73.6	60.8	58.1	65.5	76.9	65.5	60.5	**81.4**	**83.8**	**81.4**	**80.2**
	LDWBC	58.8	71.9	58.8	51.0	PBC	∪	LDWBC	63.9	77.5	63.9	58.7	68.6	77.9	68.6	63.0	**79.9**	**83.4**	**79.9**	**80.4**

Table 4
Cross-dataset performances (Accuracy - A, Precision - P, Recall - R and F1-score - F1) on the PBC dataset of the CL techniques compared to the centralised approach. The column SOURCE reports the dataset (or datasets, indicated as D1 ∪ D2) contributing to the final model. The best performance for each architecture is highlighted in boldface.

	Traditional learning					Collaborative learning													
	SOURCE	Centralised learning				SOURCE			Early fusion				Late fusion				Federated learning			
		A	P	R	F1				A	P	R	F1	A	P	R	F1	A	P	R	F1
Res.152	Raabin	40.1	50.6	40.1	32.4	Raabin	∪	AML	40.9	56.4	40.9	32.7	41.7	55.8	41.7	32.5	**50.0**	**77.7**	**50.0**	**41.4**
	AML	43.8	70.3	43.8	34.5	Raabin	∪	LDWBC	47.5	54.1	47.5	37.2	44.6	54.9	44.6	33.5	**65.3**	**78.4**	**65.3**	**61.1**
	LDWBC	52.5	54.0	52.5	44.2	AML	∪	LDWBC	49.6	78.2	49.6	39.1	47.1	56.9	47.1	36.6	**75.2**	**84.5**	**75.2**	**70.4**
Mob.V3	Raabin	55.0	55.6	55.0	48.4	Raabin	∪	AML	**80.2**	**83.0**	**80.2**	**79.8**	68.2	77.9	68.2	65.4	71.5	81.1	71.5	67.4
	AML	**71.1**	**78.7**	**71.1**	**70.1**	Raabin	∪	LDWBC	50.8	56.6	50.8	40.9	43.4	40.9	43.4	35.5	45.5	57.7	45.5	34.7
	LDWBC	33.5	47.8	33.5	25.4	AML	∪	LDWBC	**68.2**	**78.1**	**68.2**	**61.6**	62.8	77.8	62.8	58.4	55.4	73.0	55.4	46.1
Inc.v3	Raabin	47.5	47.5	47.5	39.5	Raabin	∪	AML	42.6	32.9	42.6	32.5	**53.7**	**71.0**	**53.7**	**42.7**	38.0	39.3	38.0	25.4
	AML	40.9	68.2	40.9	32.3	Raabin	∪	LDWBC	38.8	26.3	38.8	23.5	39.3	25.0	39.3	26.7	**63.6**	**57.7**	**63.6**	**56.3**
	LDWBC	25.2	15.8	25.2	14.1	AML	∪	LDWBC	24.8	17.2	24.8	13.7	31.8	19.5	31.8	20.8	**59.5**	**71.7**	**59.5**	**57.2**
Den.121	Raabin	41.3	55.6	41.3	34.2	Raabin	∪	AML	33.1	34.9	33.1	21.3	35.1	16.2	35.1	21.3	**62.8**	**79.4**	**62.8**	**59.2**
	AML	27.3	11.0	27.3	15.5	Raabin	∪	LDWBC	56.6	72.5	56.6	50.8	63.2	80.8	63.2	59.4	**69.8**	**80.2**	**69.8**	**68.7**
	LDWBC	**82.2**	**85.7**	**82.2**	**81.2**	AML	∪	LDWBC	46.7	57.9	46.7	37.9	55.0	78.4	55.0	45.4	62.0	78.6	62.0	59.3
ViT	Raabin	59.9	58.4	59.9	54.5	Raabin	∪	AML	71.1	74.7	71.1	69.2	**76.4**	**81.2**	**76.4**	**75.6**	64.0	75.5	64.0	62.7
	AML	**79.3**	**83.4**	**79.3**	**78.9**	Raabin	∪	LDWBC	72.3	78.5	72.3	69.3	69.4	82.5	69.4	66.6	70.2	80.2	70.2	68.2
	LDWBC	63.6	77.7	63.6	61.7	AML	∪	LDWBC	70.7	79.5	70.7	66.6	83.5	86.4	83.5	82.7	**84.3**	**87.2**	**84.3**	**83.6**
Swin	Raabin	60.3	61.8	60.3	54.2	Raabin	∪	AML	62.8	69.3	62.8	58.6	60.3	69.0	60.3	52.3	**74.8**	**80.6**	**74.8**	**73.5**
	AML	52.9	72.9	52.9	43.4	Raabin	∪	LDWBC	61.2	76.1	61.2	57.1	67.4	81.2	67.4	64.7	**75.2**	**81.2**	**75.2**	**74.4**
	LDWBC	71.9	82.6	71.9	71.4	AML	∪	LDWBC	54.1	74.8	54.1	48.0	59.5	78.0	59.5	55.2	**92.1**	**93.3**	**92.1**	**92.0**

outperforms both early and late fusion approaches. Across the majority of evaluations, FL exhibits superior performance compared to the other two DCML methods, showcasing enhanced generalisation and robustness against variations in datasets.

5.2. Qualitative analysis

To conduct a qualitative assessment of the compared methods, we engaged in a visual examination of the quality of features extracted from the intermediate layers of the models, as well as the resulting activation maps. For brevity, our analysis focuses exclusively on the results obtained with Res.152, which exhibited superior performance under the conventional centralised approach. This selection allows for more precise insights into the advantages offered by the DCML paradigm. Notably, we exclude the Late Fusion approach from our analysis, as it involves fusion at the final layer, resulting in data extracted from the intermediate layers identical to those of the centralised approach.

Initially, we explore the distribution of features through t-SNE [90], aiming to assess the models' capability in clustering cells of the same class while simultaneously separating cells of different classes. The plot in Fig. 4 includes features from various target datasets to gauge generalisation capabilities. Markers of the same colours forming distinct groups represent cells of the same class with notable similarity in embedding space. Our observations indicate that while the centralised approach achieves high intra-class clustering, it fails to introduce meaningful separation between cells of different classes. The early fusion technique exhibits more meaningful separation, albeit with scattered records and less well-defined clusters of cells of the same class. Conversely, the FL technique strikes a better balance between intra-class clustering and inter-class separability.

Additionally, to offer a visual explanation for the decisions made by the DL models and enhance transparency, we extracted the Class Activation Map (CAMs) from each dataset, again using the Res.152 architecture, which, as mentioned before, exhibited superior performance under the conventional centralised approach. Given the clinical

Table 5
Cross-dataset performances (Accuracy - A, Precision - P, Recall - R and F1-score - F1) on the LDWBC dataset of the CL techniques compared to the centralised approach. The column SOURCE reports the dataset (or datasets, indicated as D1 ∪ D2) contributing to the final model. The best performance for each architecture is highlighted in boldface.

	Traditional learning					Collaborative learning													
	SOURCE	Centralised learning				SOURCE			Early fusion				Late fusion				Federated learning			
		A	P	R	F1				A	P	R	F1	A	P	R	F1	A	P	R	F1
Res.152	Raabin	53.3	55.5	53.3	49.0	Raabin	∪	AML	52.9	54.8	52.9	45.2	**56.4**	53.8	**56.4**	**50.6**	52.0	**57.6**	52.0	47.6
	AML	50.7	52.3	50.7	45.8	Raabin	∪	PBC	72.0	78.5	72.0	71.3	79.1	**82.0**	79.1	79.1	**80.9**	**82.0**	**80.9**	**80.7**
	PBC	**79.1**	81.1	**79.1**	**78.2**	AML	∪	PBC	70.2	79.5	70.2	67.3	72.4	**82.0**	72.4	71.4	59.6	75.5	59.6	57.7
Mob.V3	Raabin	47.6	62.1	47.6	46.3	Raabin	∪	AML	57.3	50.2	57.3	51.6	56.9	57.1	56.9	52.4	**61.8**	**67.1**	**61.8**	**60.2**
	AML	39.1	50.6	39.1	33.0	Raabin	∪	PBC	52.4	57.5	52.4	46.9	56.0	63.9	56.0	50.2	**64.4**	**67.7**	**64.4**	**62.7**
	PBC	57.8	71.9	57.8	52.8	AML	∪	PBC	66.7	71.8	66.7	65.3	**67.6**	**73.5**	**67.6**	**66.6**	60.9	68.2	60.9	56.5
Inc.v3	Raabin	**51.6**	60.4	**51.6**	**52.1**	Raabin	∪	AML	50.7	**65.2**	50.7	44.6	49.3	63.5	49.3	48.2	48.9	56.9	48.9	46.0
	AML	45.8	50.7	45.8	41.8	Raabin	∪	PBC	59.6	69.3	59.6	57.0	63.1	67.5	63.1	63.2	**68.0**	**72.2**	**68.0**	**66.9**
	PBC	59.1	61.7	59.1	59.0	AML	∪	PBC	58.7	50.3	58.7	52.7	57.3	64.8	57.3	56.6	**63.6**	**74.8**	**63.6**	**63.3**
Den.121	Raabin	45.3	62.6	45.3	41.0	Raabin	∪	AML	40.9	40.2	40.9	33.4	43.1	50.3	43.1	37.1	**61.3**	**67.7**	**61.3**	**61.0**
	AML	40.9	38.2	40.9	33.0	Raabin	∪	PBC	57.8	73.0	57.8	56.1	**60.4**	**76.0**	**60.4**	**59.8**	56.0	73.8	56.0	55.1
	PBC	61.3	62.2	61.3	59.4	AML	∪	PBC	**65.8**	71.6	**65.8**	63.1	59.6	73.5	59.6	57.4	65.3	**77.9**	65.3	**63.4**
ViT	Raabin	55.1	66.3	55.1	53.3	Raabin	∪	AML	51.1	55.5	51.1	47.8	**60.9**	**71.7**	**60.9**	**60.0**	49.8	68.9	49.8	44.9
	AML	62.7	75.1	62.7	61.9	Raabin	∪	PBC	62.2	69.2	62.2	60.9	**65.8**	**74.2**	**65.8**	**65.1**	63.6	72.6	63.6	63.9
	PBC	66.2	69.9	66.2	65.2	AML	∪	PBC	56.0	62.8	56.0	52.9	**68.0**	**77.2**	**68.0**	**68.0**	62.7	74.0	62.7	61.5
Swin	Raabin	52.9	56.5	52.9	50.2	Raabin	∪	AML	50.7	59.6	50.7	44.2	52.9	56.9	52.9	46.1	**69.3**	**73.2**	**69.3**	**64.2**
	AML	44.9	40.5	44.9	35.8	Raabin	∪	PBC	52.4	60.4	52.4	46.7	58.7	63.7	58.7	55.4	**62.7**	**67.1**	**62.7**	**61.1**
	PBC	60.4	71.2	60.4	57.0	AML	∪	PBC	48.9	70.7	48.9	37.0	50.7	65.9	50.7	42.1	**71.1**	**76.9**	**71.1**	**70.7**

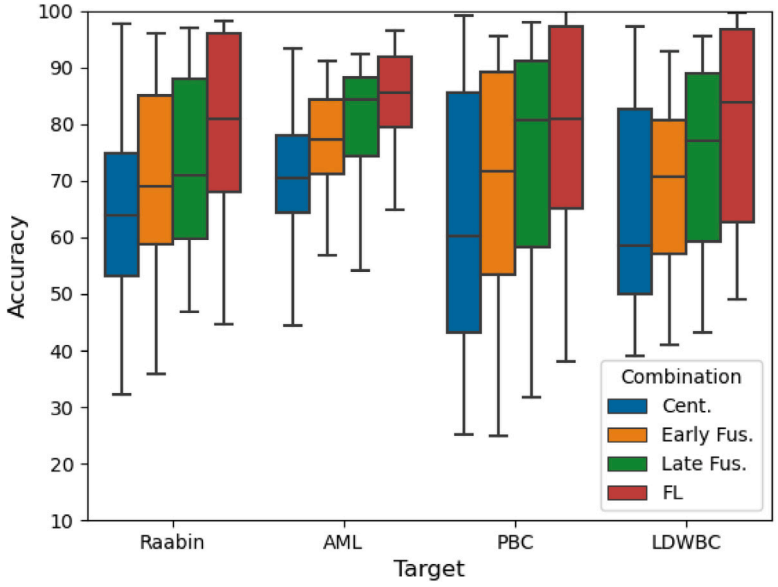

Fig. 3. Analysis of accuracy trend of the DCML techniques, compared to the centralised approach, on different target datasets. For brevity, we have reported the average performance of all DL architectures trained with the same SOURCE dataset (or datasets) reported in Tables 2–5.

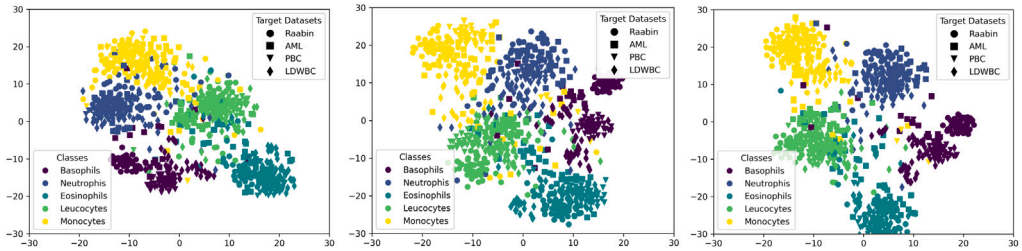

Fig. 4. The two-dimensional projection obtained with t-SNE of the high-dimensional feature vectors extracted from the four target datasets by using the first convolutional layer of the fourth residual block of the Res.152 model obtained with the centralised approach (left), early fusion (centre) and FL (right).

importance of relying on interpretable cues, such as cytoplasmic texture, nuclear shape, and granularity, we expect effective DL models to localise such regions in WBC images when making decisions.

We employed two algorithms to extract the CAMs, namely the Gradient-weighted CAM (Grad-CAM) algorithm [91] and the Eigen-CAM [92]. Both methods aim to highlight the regions in the input

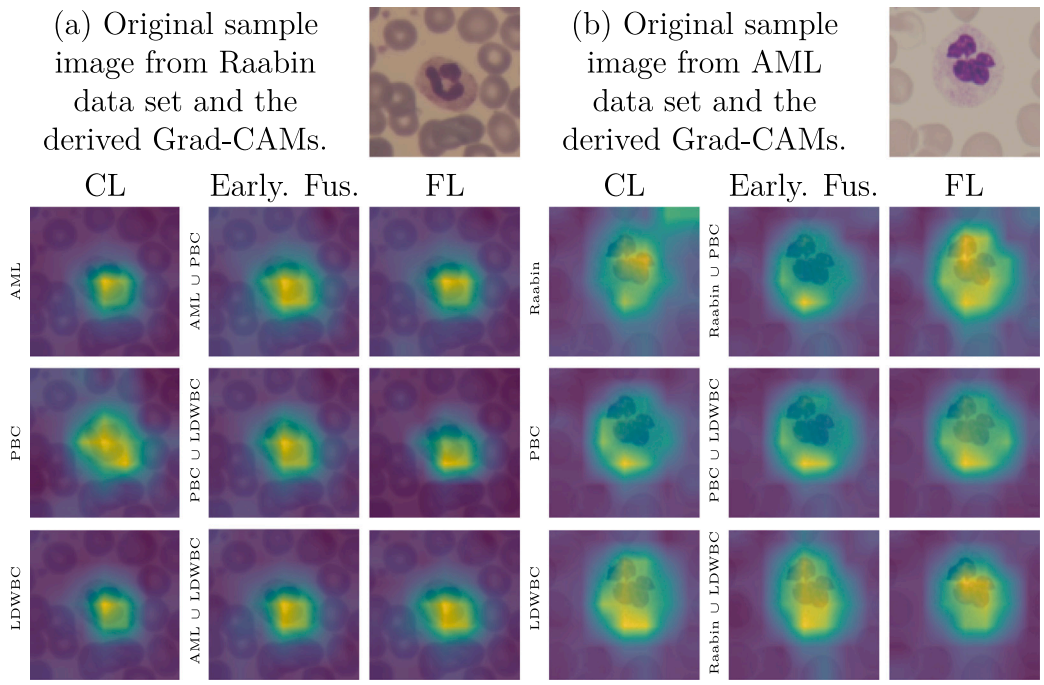

Fig. 5. Comparison of the activation maps, computed with the Grad-CAM algorithm, on Raabin (a) and AML (b) datasets, from the Res.152 model trained with the dataset (or datasets, indicated as D1 ∪ D2) reported on the left, using the centralised learning (CL), early fusion (Early Fus.) and federated learning (FL) approaches.

image that most influence the model's predictions, thus helping to assess whether predictions are grounded in relevant morphological and textural features of WBCs. Grad-CAM [91] is a gradient-based explainability technique that generates coarse localisation heatmaps by leveraging the gradients of the target class flowing into the final convolutional layer. These gradients are globally averaged and used as weights to produce a weighted combination of activation maps, highlighting the regions most relevant for a given prediction. On the contrary, Eigen-CAM [92] is a gradient-free alternative that constructs activation maps by performing Principal Component Analysis (PCA) over the activation tensors of convolutional layers. Specifically, the first principal component is projected back into the image space to form a saliency map. This method provides a stable, architecture-agnostic visualisation that does not rely on backpropagation, making it particularly useful when gradient information is noisy or unavailable.

Even in this context, focusing on generalisation capabilities, we present cross-dataset results in Figs. 5 and 7 and Figs. 6 and 8. Each row displays activation maps of the same DL models trained with a source dataset or pair of datasets and tested with each of the aforementioned datasets. Contrary to expectations, DL models associated with the centralised approach, particularly evident in cross-dataset results with Grad-CAM, exhibit activations not solely focused on WBCs. However, early fusion and FL approaches alleviate this issue, with minimal inaccuracies observed in the activation maps. Indeed, Figs. 5 and 7 show that the only imperfections consist of the maximum activation values not centred in the WBCs. These inaccuracies are minimal and, notably, FL demonstrates consistent activation map quality with both Grad-CAM and Eigen-CAM algorithms.

Compared to Grad-CAM, Eigen-CAM produced more visually stable and coherent activation patterns across domains and collaborative setups. However, while Grad-CAM more clearly highlighted differences between learning strategies (centralised, early fusion, and FL), Eigen-CAM maps were more uniform, indicating that despite the training differences, the models consistently focused on morphologically relevant regions, neglecting irrelevant or artefactual features.

5.3. Comparison with related work

It is imperative to highlight that this comparative analysis does not aim to propose a specialised neural network for WBC subtypes classification; rather, it focuses on enhancing the generalisation of existing (including ad hoc) architectures when confronted with domain shift.

In this context, very few studies have conducted cross-dataset experiments on heterogeneous datasets to evaluate the generalisation capacity of their methodologies [15]; also, some methods have different prerogatives [14,38,54,55].

In the study conducted by Salehi et al. [15], the authors proposed a novel approach for extracting features from WBCs in PBS, based on a cross-domain adapted autoencoder to address domain shifts, called AutoEncoder-based Cell Feature Extractor (AE-CFE). After extracting features in an unsupervised manner, the approach was trained on AML and tested on PBC, and vice versa. The method obtained performance of 21.90% and 45.10% in terms of accuracy, resulting in lower accuracy than the used approach.

Among the works with different objectives, Tummala et al. [38] proposed a Siamese twin network (STN) model using EfficientNet-B3 as the base model for WBC classification in various few-shot scenarios. Additionally, Sadafi et al. proposed a continual learning approach for class incremental and domain incremental scenarios in WBC classification, involving the merging of two public datasets, PBS and AML, along with a private one [54].

For these reasons, in Table 6 we present a comprehensive summary of existing research results on the same datasets considered in this study, outlining their methodologies and corresponding accuracies.

Datasets results. In the PBC dataset, a notable spectrum of accuracies is observed among the methodologies. Particularly, the feature extraction technique known as LeuFeatx, paired with the XGBoost classifier, yields the lowest accuracy of 92.00% [32]. Conversely, the ML-CNN [31] achieves the highest accuracy of 99.72%, surpassing all other approaches examined in this dataset.

In the case of the Raabin dataset, the ML-CNN similarly exhibits exceptional performance, attaining an accuracy of 99.22% [31], closely

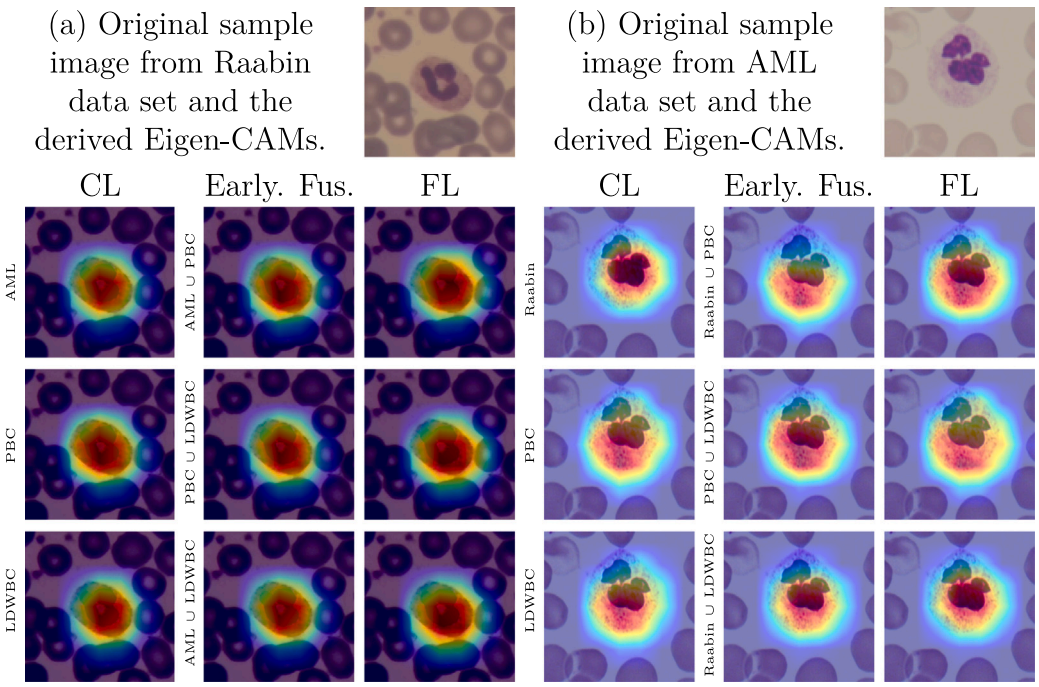

Fig. 6. Comparison of the activation maps, computed with the Eigen-CAM algorithm, on Raabin (a) and AML (b) datasets, from the Res.152 model trained with the dataset (or datasets, indicated as D1 ∪ D2) reported on the left, using the centralised learning (CL), early fusion (Early Fus.) and federated learning (FL) approaches.

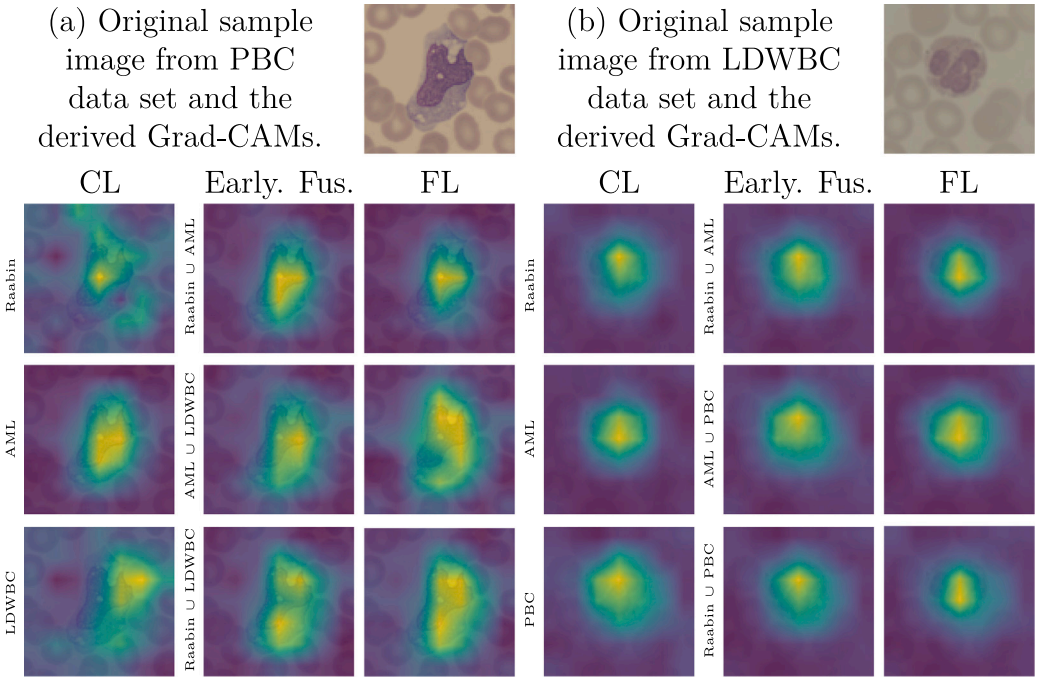

Fig. 7. Comparison of the activation maps, computed with the Grad-CAM algorithm, on PBC (a) and LDWBC (b) datasets, from the Res.152 model trained with the dataset (or datasets, indicated as D1 ∪ D2) reported on the left, using the centralised learning (CL), early fusion (Early Fus.) and federated learning (FL) approaches.

trailed by ResNext50 [5] with 99.17%. These findings underscore the efficacy of tailored CNNs in WBC classification tasks, even though the method with the least accuracy on this dataset is WBCaps, achieving 92.74%.

To our knowledge, only one method has been tested on the LDWBC dataset [75]. This method amalgamates ResNet-50 and DenseNet-121, achieving an accuracy of 97.84%.

Moving to the AML dataset, Ghost-ResNeXt attains the highest accuracy of 98.61% [43], closely followed by DCNN with 98.27% [93]. Conversely, the lowest accuracy on this dataset is recorded with the DCAE method, achieving 93.12% [94].

Generalisation results discussion. Transitioning to a broader discussion, as emphasised by Table 6 it is pertinent to underscore that on the one hand the various methodologies within this domain are specifically tailored to represent and classify WBCs. On the other hand, only one work in this domain conducted a cross-dataset evaluation, a critical aspect for assessing the generalisation capabilities of methodologies.

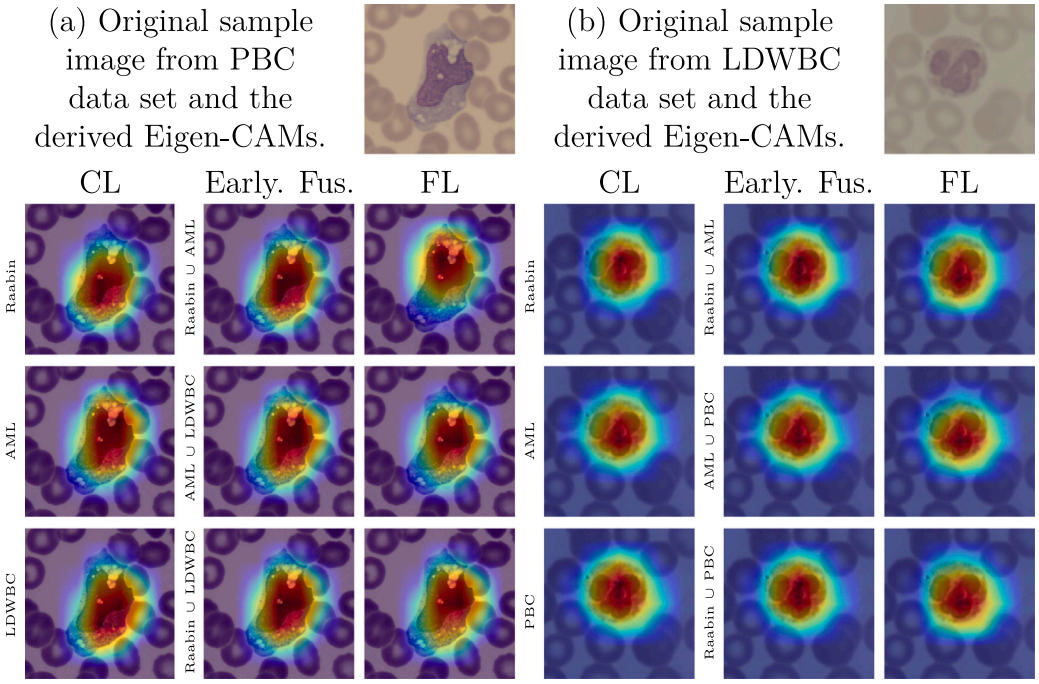

Fig. 8. Comparison of the activation maps, computed with the Eigen-CAM algorithm, on PBC (a) and LDWBC (b) datasets, from the Res.152 model trained with the dataset (or datasets, indicated as D1 ∪ D2) reported on the left, using the centralised learning (CL), early fusion (Early Fus.) and federated learning (FL) approaches.

Table 6
Overview of existing works in this context with key insights from the proposed methods.

DS	Method	Year	Accuracy (%)
PBC	VGG-16 [30]	2019	96.20
	ShuffleNet [95]	2020	97.94
	BloodCaps (Capsule Network) [37]	2021	99.30
	ResNext [15]	2022	85.70
	LeuFeatx + XGBoost [32]	2022	92.00
	EfficientNetV2 B0 [96]	2023	93.00
	MedVIT-L [41]	2023	95.40
	FeSViBS [40]	2023	93.60
	Siamese twin network (STN) [38]	2023	97.25
	EnsDA_Mix (with DenseNet-201) [42]	2023	99.47
	RCMNet [39]	2023	99.63
	Multibranch Lightweight CNN (ML-CNN) [31]	2024	99.72
Raabin	WBCaps [97]	2020	92.74
	NucSegNet [98]	2020	93.08
	Shape and Color features + SVM [33]	2021	94.65
	CNNDC [34]	2021	93.41
	ResNext50 [5]	2022	99.17
	ResNet50 + DenseNet121 [75]	2022	98.71
	DRFA-Net [36]	2022	95.17
	DeepViT [9]	2023	97.00
	Meta-Learning Methodology (MeLM) [99]	2023	98.29
	VGG-16 [12]	2023	98.75
	Color invariance + ResNet50 [44]	2024	97.80
	ML-CNN [31]	2024	99.22
LDWBC	ResNet50 + DenseNet121 [75]	2022	97.84
AML	Random forest [100]	2020	93.40
	Deep Convolutional Neural network (DCNN) [93]	2021	98.27
	ResNext [28]	2022	96.10
	XGBoost [101]	2022	97.57
	Ghost-ResNeXt [43]	2023	98.61
	Deep convolutional autoencoder (DCAE) [94]	2023	93.12
	Model and Data Heterogeneity in FL (MDH-FL) [63]	2024	81.69

In particular, Tsutsui et al. [12] focuses on assessing the performance of WBCs classification under domain shift. They achieve this objective by training two established CNNs, namely ResNet-50 and VGG-16, on the Raabin dataset and subsequently evaluating their performance on both Raabin and the LISC dataset. The study highlights the challenge of poor generalisation exhibited by the ResNet classifier when

faced with domain shifts, as evidenced by the model's inferior performance when trained on Raabin and tested on LISC. This is attributed to the utilisation of batch normalisation in ResNet. To address this limitation, the authors propose the adoption of group normalisation, which yields results comparable to those obtained with VGG-16, a model devoid of batch normalisation layers.

As previously mentioned in Section 2, we identified two studies that applied FL methodologies for WBC classification [63,64]. Both studies focused their evaluation on single datasets and employed non-IID strategies to partition the data among different clients. However, they did not explore cross-dataset scenarios or account for domain shifts. The studies aimed to classify known classes and identify unknown classes separately, whereas our current investigation focuses on recognising five distinct WBC subtypes across different datasets while addressing domain shifts.

Without aiming for a direct point-by-point comparison, it is noteworthy that our study achieves accuracies comparable to local (non-FL) methods. This achievement is significant as it allows for privacy preservation while relating multiple clients with heterogeneous datasets, thus highlighting the potential of our approach in real-world scenarios where data privacy [19,20] and diverse data sources are paramount concerns.

Some examples in this context are given by the authors in [9], where they addressed WBC classification utilising the Raabin dataset and the DeepViT neural network, an enhanced iteration of the ViT network utilised in this investigation. Nonetheless, their methodology did not undergo evaluation through cross-dataset validation. Furthermore, our experiments indicate that when this neural network is solely trained on the Raabin dataset, its performance substantially declines to below 60% across all metrics considered (see Tables 2–5).

Similarly, in [75], the authors integrated the strengths of ResNet and DenseNet modules to enhance feature representation and incorporated a spatial and channel attention module (SCAM) to augment the model's discriminative prowess for WBC nucleus characterisation. They evaluated their approach on LDWBC and Raabin datasets, achieving respective accuracy scores of 97.84% and 98.71%, precision values of 91.61% and 97.18%, recall scores of 96.38% and 98.42%, and F1-scores of 93.82% and 97.78%. Regrettably, the authors did not subject their approach to cross-dataset assessment. Nevertheless, given their method's adept fusion of DenseNet and ResNet architectures tailored for WBC classification, it facilitates a comparative analysis with our findings. As delineated in Tables 2–5, the results derived from the consolidated DenseNet and ResNet architectures fall short of those reported in the aforementioned study. However, as underscored by Tables 2–5, a centralised approach reduces neural networks' abstraction capabilities, rendering them impractical for adaptation to alternate source datasets or validation with new WBC images. Consequently, FL and DCML methods could potentially enhance the applicability and utility of previously proposed neural networks in real-world scenarios, catering to diverse medical centres and assisting biologists in WBC classification endeavours.

5.4. Discussion

It appears from the above quantitative and qualitative results that FL stands out from the other approaches. It should also be noted that this average increase in performance corresponds to an increase in costs, not only related to computational costs but also to communications and security, even when FedAVG is used as an optimisation method. Indeed, it requires the establishment of prolonged private and secure communication between the clients and the server, which are responsible for training and data aggregation/fusion, respectively.

During the training of the models with a centralised approach, we found that the timings vary from an average (there are variabilities due to the size of the dataset used) of 38 m for Mob.v3, which is the lightest architecture to 8h2 m for Swin, which is the heaviest architecture. It should be remembered that these timelines also apply to early fusion and late fusion approaches and simply need to be multiplied by n, i.e., the number of haematological research centres involved. Indeed, for these DCML approaches, there is no continuous communication between different haematological research centres, but it can be assumed that a single communication is established to consolidate a single model (in the case of early fusion) or to exchange the weights of the other $n-1$ branches that make up the final model (in the case of late fusion).

When training the models with FL approach, we found that the timings vary from an average of 4h23 m for Mob.v3 to 14h22 m for Swin. These timelines indicate that the communication overhead greatly influences the overall timings for training. Considering what also emerged regarding the vulnerabilities related to the possibilities of reconstructing (even if partially) the training data [23,24], it becomes imperative to ensure the safeguarding of all nodes, fostering secure communication among them, which also increases the material costs for this purpose.

In addition, it should be mentioned that all approaches except late fusion result in a single consolidated model. This means that for this DCML approach, in which a new model with multiple branches is created, there are additional costs in the inference phase, proportional to the number of branches themselves.

Finally, in our analysis, we acknowledge that we opted for a value of $n = 2$ due to constraints pertaining to time and resources. However, while increasing the number of nodes undergoing training would possibly prolong the training time, it is plausible to believe that greater involvement of blood research centres, thus expanding the number of datasets, could potentially further improve the generalisation capabilities of the DCML system.

6. Conclusion

In this work, we revisited the DCML paradigm approach for WBC subtypes classification, originally proposed with different purposes, arguing that it can be a further effective solution against domain shift, besides existing solutions based on domain adaptation (either supervised or unsupervised) and domain generalisation. Under this viewpoint, DCML can be viewed as a domain generalisation approach which leverages the collaboration of different data sources in a privacy-preserving setup. An extensive empirical evaluation of three well-known DCML methods, the early and late fusion and FL approaches, on four benchmark WBC datasets, confirmed the effectiveness of the DCML paradigm, which turned out to outperform the classical centralised approach. Considering also that the proposed solution against domain shift is very general and model-agnostic, and can therefore be implemented on top of *any* classification system, its potential scope is broader than the one considered in this work: (i) Besides being an alternative to other methods against domain shift, it is also *complementary* to them and can be combined with them to improve their performance further; (ii) Similarly, it can also be used in application scenarios where the target and source domains coincide to improve the model's effectiveness further; (iii) Finally, it can also be used in further medical applications affected by restrictive privacy-preserving policies like the one addressed in this work or even other non-medical applications (e.g., video surveillance). In this context, an interesting research direction for future work is to investigate the combination of the DCML paradigm and other methods against domain shift and assess DCML's effectiveness in different medical applications or even tasks, i.e., image segmentation.

CRediT authorship contribution statement

Lorenzo Putzu: Writing – review & editing, Supervision, Project administration, Investigation, Data curation, Visualization, Writing – original draft, Validation, Resources, Formal analysis, Software, Methodology, Conceptualization. **Simone Porcu:** Visualization, Supervision,

Investigation, Writing – review & editing, Validation, Methodology, Formal analysis, Conceptualization, Writing – original draft. **Andrea Loddo:** Visualization, Methodology, Writing – original draft, Supervision, Formal analysis, Writing – review & editing, Validation, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was partially supported by the project SERICS (PE00000 014) under the NRRP MUR program funded by the EUNGEU - CUP F53C22000740007.

Data availability

The authors do not have permission to share data.

References

[1] H. Kutlu, E. Avci, F. Özyurt, White blood cells detection and classification based on regional convolutional neural networks, Med. Hypotheses 135 (2020) 109472.

[2] H. Nohynek, E. Valkeila, M. Leinonen, Erythrocyte sedimentation rate, white blood cell count and serum C-reactive protein in assessing etiologic diagnosis of acute lower respiratory infections in children, Pediatr. Infect. Dis. J. 14 (6) (1995) 484–489.

[3] A. Veronelli, M. Laneri, R. Ranieri, D. Koprivec, D. Vardaro, M. Paganelli, F. Folli, A.E. Pontiroli, White blood cells in obesity and diabetes: effects of weight loss and normalization of glucose metabolism, Diabetes Care 27 (10) (2004) 2501–2502.

[4] A.G. Burton, K.E. Jandrey, Leukocytosis and Leukopenia, Textb. Small Anim. Emerg. Med. (2018) 405–412.

[5] Z.M. Kouzehkanan, S. Saghari, S. Tavakoli, P. Rostami, M. Abaszadeh, F. Mirzadeh, E.S. Satlsar, M. Gheidishahran, F. Gorgi, S. Mohammadi, et al., A large dataset of white blood cells containing cell locations and types, along with segmented nuclei and cytoplasm, Sci. Rep. 12 (1) (2022) 1123.

[6] L. Putzu, G. Caocci, C. Di Ruberto, Leucocyte classification for leukaemia detection using image processing techniques, Artif. Intell. Med. 62 (3) (2014) 179–191.

[7] R.L. Siegel, K.D. Miller, N.S. Wagle, A. Jemal, Cancer statistics, 2023, Ca Cancer J. Clin. 73 (1) (2023) 17–48.

[8] S. Saleem, J. Amin, M. Sharif, G.A. Mallah, S. Kadry, A.H. Gandomi, Leukemia segmentation and classification: A comprehensive survey, Comput. Biol. Med. 150 (2022) 106028.

[9] R. Rubin, S.M. Anzar, A. Panthakkan, W. Mansoor, Transforming healthcare: Raabin white blood cell classification with deep vision transformer, in: 2023 6th International Conference on Signal Processing and Information Security, ICSPIS, 2023, pp. 212–217, http://dx.doi.org/10.1109/ICSPIS60075.2023.10344258.

[10] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J.N. Chiang, Z. Wu, X. Ding, Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation, Med. Image Anal. 63 (2020) 101693.

[11] A. Loddo, L. Putzu, On the effectiveness of leukocytes classification methods in a real application scenario, AI 2 (3) (2021) 394–412, http://dx.doi.org/10.3390/ai2030025, URL https://www.mdpi.com/2673-2688/2/3/25.

[12] S. Tsutsui, Z. Su, B. Wen, Benchmarking white blood cell classification under domain shift, in: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, Institute of Electrical and Electronics Engineers Inc. 2023, http://dx.doi.org/10.1109/ICASSP49357.2023.10097167.

[13] H. Chen, J. Liu, C. Hua, Z. Zuo, J. Feng, B. Pang, D. Xiao, TransMixNet: An attention based double-branch model for white blood cell classification and its training with the fuzzified training data, in: Proceedings - 2021 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2021, Institute of Electrical and Electronics Engineers Inc. 2021, pp. 842–847.

[14] P. Pandey, P.A. P., V. Kyatham, D. Mishra, T.R. Dastidar, Target-independent domain adaptation for WBC classification using generative latent search, IEEE Trans. Med. Imag. 39 (12) (2020) 3979–3991.

[15] R. Salehi, A. Sadafi, A. Gruber, P. Lienemann, N. Navab, S. Albarqouni, C. Marr, Unsupervised cross-domain feature extraction for single blood cell image classification, in: L. Wang, Q. Dou, P.T. Fletcher, S. Speidel, S. Li (Eds.), Medical Image Computing and Computer Assisted Intervention - MICCAI 2022 - 25th International Conference, Singapore, September 18-22, 2022, Proceedings, Part III, in: Lecture Notes in Computer Science, vol. 13433, Springer, 2022, pp. 739–748.

[16] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, P.S. Yu, Generalizing to unseen domains: A survey on domain generalization, IEEE Trans. Knowl. Data Eng. 35 (8) (2023) 8052–8072.

[17] B. Felfeliyan, N.D. Forkert, A.R. Hareendranathan, D. Cornel, Y. Zhou, G. Kuntze, J.L. Jaremko, J.L. Ronsky, Self-supervised-RCNN for medical image segmentation with limited data annotation, Comput. Med. Imag. Graph. 109 (2023) 102297.

[18] Q. Yang, Y. Liu, T. Chen, Y. Tong, Federated machine learning: Concept and applications, ACM Trans. Intell. Syst. Technol. 10 (2) (2019) http://dx.doi.org/10.1145/3298981.

[19] V. Mothukuri, R.M. Parizi, S. Pouriyeh, Y. Huang, A. Dehghantanha, G. Srivastava, A survey on security and privacy of federated learning, Future Gener. Comput. Syst. 115 (2021) 619–640, http://dx.doi.org/10.1016/j.future.2020.10.007, URL https://www.sciencedirect.com/science/article/pii/S0167739X20329848.

[20] M. Narula, J. Meena, D.K. Vishwakarma, A comprehensive review on federated learning for data-sensitive application: Open issues & challenges, Eng. Appl. Artif. Intell. 133 (2024) 108128, http://dx.doi.org/10.1016/j.engappai.2024.108128.

[21] P. Kairouz, H.B. McMahan, et al., Advances and open problems in federated learning, Found. Trends Mach. Learn. 14 (1–2) (2021) 1–210.

[22] P. Kairouz, et al., Advances and open problems in federated learning, Found. Trends® Mach. Learn. 14 (1–2) (2021) 1–210, http://dx.doi.org/10.1561/2200000083.

[23] F. Boenisch, A. Dziedzic, R. Schuster, A.S. Shamsabadi, I. Shumailov, N. Papernot, When the curious abandon honesty: Federated learning is not private, in: Proceedings - 8th IEEE European Symposium on Security and Privacy, Euro S and P 2023, Institute of Electrical and Electronics Engineers Inc. 2023, pp. 175–199, http://dx.doi.org/10.1109/EUROSP57164.2023.00020.

[24] Q. Wang, D. Kurz, Reconstructing training data from diverse ML models by ensemble inversion, in: Proceedings - 2022 IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Institute of Electrical and Electronics Engineers Inc. 2022, pp. 3870–3878, http://dx.doi.org/10.1109/WACV51458.2022.00392.

[25] S. Boulahia, A. Amamra, M. Madi, S. Daikh, Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition, Mach. Vis. Appl. 32 (2021) http://dx.doi.org/10.1007/s00138-021-01249-8.

[26] A. Acevedo, A. Merino, L. Boldú, A. Molina, S. Alférez, J. Rodellar, A new convolutional neural network predictive model for the automatic recognition of hypogranulated neutrophils in myelodysplastic syndromes, Comput. Biol. Med. 134 (2021) 104479.

[27] Q. Huang, W. Li, B. Zhang, Q. Li, R. Tao, N.H. Lovell, Blood cell classification based on hyperspectral imaging with modulated gabor and CNN, IEEE J. Biomed. Heal. Informatics 24 (1) (2020) 160–170.

[28] C. Matek, S. Schwarz, K. Spiekermann, C. Marr, Human-level recognition of blast cells in acute myeloid leukaemia with convolutional neural networks, Nat. Mach. Intell. 1 (11) (2019) 538–544.

[29] L.H. Vogado, R.M. Veras, F.H. Araujo, R.R. Silva, K.R. Aires, Leukemia diagnosis in blood slides using transfer learning in CNNs and SVM for classification, Eng. Appl. Artif. Intell. 72 (2018) 415–422.

[30] A. Acevedo, S. Alférez, A. Merino, L. Puigví, J. Rodellar, Recognition of peripheral blood cell images using convolutional neural networks, Comput. Methods Programs Biomed. 180 (2019) 105020.

[31] H. Fırat, Classification of microscopic peripheral blood cell images using multibranch lightweight CNN-based model, Neural Comput. Appl. 36 (4) (2024) 1599–1620.

[32] P. Rastogi, K. Khanna, V. Singh, LeuFeatx: Deep learning-based feature extractor for the diagnosis of acute leukemia from microscopic images of peripheral blood smear, Comput. Biol. Med. 142 (2022) 105236.

[33] S. Tavakoli, A. Ghaffari, Z.M. Kouzehkanan, R. Hosseini, New segmentation and feature extraction algorithm for classification of white blood cells in peripheral smear images, Sci. Rep. 11 (1) (2021) 19428.

[34] P.K. Das, S. Meher, An efficient deep convolutional neural network based detection and classification of acute lymphoblastic Leukemia, Expert Syst. Appl. 183 (2021) 115311.

[35] A. Şengür, Y. Akbulut, Ü. Budak, Z. Cömert, White blood cell classification based on shape and deep features, in: 2019 International Artificial Intelligence and Data Processing Symposium, IDAP, IEEE, 2019, pp. 1–4.

[36] L. Jiang, C. Tang, H. Zhou, White blood cell classification via a discriminative region detection assisted feature aggregation network, Biomed. Opt. Express 13 (10) (2022) 5246–5260.

[37] F. Long, J. Peng, W. Song, X. Xia, J. Sang, BloodCaps: A capsule network based model for the multiclassification of human peripheral blood cells, Comput. Methods Programs Biomed. 202 (2021) 105972.

[38] S. Tummala, A.K. Suresh, Few-shot learning using explainable Siamese twin network for the automated classification of blood cells, Med. Biol. Eng. Comput. 61 (6) (2023) 1549–1563.

[39] R. Zhang, X. Han, Z. Lei, C. Jiang, I. Gul, Q. Hu, S. Zhai, H. Liu, L. Lian, Y. Liu, Y. Zhang, Y. Dong, C.Y. Zhang, T.K. Lam, Y. Han, D. Yu, J. Zhou, P. Qin, RCMNet: A deep learning model assists CAR-T therapy for leukemia, Comput. Biol. Med. 150 (2022) 106084.

[40] F. Almalik, N. Alkhunaizi, I. Almakky, K. Nandakumar, FeSViBS: Federated split learning of vision transformer with block sampling, in: H. Greenspan, A. Madabhushi, P. Mousavi, S. Salcudean, J. Duncan, T.F. Syeda-Mahmood, R.H. Taylor (Eds.), Medical Image Computing and Computer Assisted Intervention - MICCAI 2023 - 26th International Conference, Vancouver, BC, Canada, October 8-12, 2023, Proceedings, Part II, in: Lecture Notes in Computer Science, vol. 14221, Springer, 2023, pp. 350–360.

[41] O.N. Manzari, H. Ahmadabadi, H. Kashiani, S.B. Shokouhi, A. Ayatollahi, MedViT: a robust vision transformer for generalized medical image classification, Comput. Biol. Med. 157 (2023) 106791.

[42] R. Bravin, L. Nanni, A. Loreggia, S. Brahnam, M. Paci, Varied image data augmentation methods for building ensemble, IEEE Access 11 (2023) 8810–8823.

[43] S.S.R. Bairaboina, S.R. Battula, Ghost-ResNeXt: An effective deep learning based on mature and immature WBC classification, Appl. Sci. 13 (6) (2023) 4054.

[44] C. Li, Y. Liu, Improved generalization of white blood cell classification by learnable illumination intensity invariant layer, IEEE Signal Process. Lett. 31 (2024) 176–180.

[45] M. Togaçar, B. Ergen, Z. Cömert, Classification of white blood cells using deep features obtained from convolutional neural network models based on the combination of feature selection methods, Appl. Soft Comput. 97 (Part B) (2020) 106810.

[46] C. Li, X. Lin, Y. Mao, W. Lin, Q. Qi, X. Ding, Y. Huang, D. Liang, Y. Yu, Domain generalization on medical imaging classification using episodic training with task augmentation, Comput. Biol. Med. 141 (2022) 105144.

[47] J. Roels, J. Hennies, Y. Saeys, W. Philips, A. Kreshuk, Domain adaptive segmentation in volume electron microscopy imaging, in: 16th IEEE International Symposium on Biomedical Imaging, ISBI 2019, Venice, Italy, April 8-11, 2019, IEEE, 2019, pp. 1519–1522.

[48] F. Mahmood, R.J. Chen, N.J. Durr, Unsupervised reverse domain adaptation for synthetic medical images via adversarial training, IEEE Trans. Med. Imag. 37 (12) (2018) 2572–2581.

[49] T. Miller, J. Cheng, H. Fu, Z. Gu, Y. Xiao, K. Zhou, S. Gao, R. Zheng, J. Liu, Correction to "noise adaptation generative adversarial network for medical image analysis", IEEE Trans. Med. Imag. 39 (7) (2020) 2566–2567.

[50] W. Li, D. Yang, C. Ma, L. Liu, Identifying novel disease categories through divergence optimization: An approach to prevent misdiagnosis in medical imaging, Comput. Biol. Med. 165 (2023) 107403.

[51] Z. Chen, Y. Pan, Y. Ye, H. Cui, Y. Xia, Treasure in distribution: A domain randomization based multi-source domain generalization for 2D medical image segmentation, in: H. Greenspan, A. Madabhushi, P. Mousavi, S. Salcudean, J. Duncan, T.F. Syeda-Mahmood, R.H. Taylor (Eds.), Medical Image Computing and Computer Assisted Intervention - MICCAI 2023 - 26th International Conference, Vancouver, BC, Canada, October 8-12, 2023, Proceedings, Part IV, in: Lecture Notes in Computer Science, vol. 14223, Springer, 2023, pp. 89–99.

[52] H. Li, S.J. Pan, S. Wang, A.C. Kot, Domain generalization with adversarial feature learning, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, Computer Vision Foundation / IEEE Computer Society, 2018, pp. 5400–5409.

[53] Q. Liu, C. Chen, J. Qin, Q. Dou, P. Heng, FedDG: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, Virtual, June 19-25, 2021, Computer Vision Foundation / IEEE, 2021, pp. 1013–1023.

[54] A. Sadafi, R. Salehi, A. Gruber, S.S. Boushehri, P. Giehr, N. Navab, C. Marr, A continual learning approach for cross-domain white blood cell classification, in: L.M. Koch, M.J. Cardoso, E. Ferrante, K. Kamnitsas, M. Islam, M. Jiang, N. Rieke, S.A. Tsaftaris, D. Yang (Eds.), Domain Adaptation and Representation Transfer - 5th MICCAI Workshop, DART 2023, Held in Conjunction with MICCAI 2023, Vancouver, BC, Canada, October 12, 2023, Proceedings, in: Lecture Notes in Computer Science, vol. 14293, Springer, 2023, pp. 136–146.

[55] R.M. Umer, A. Gruber, S.S. Boushehri, C. Metak, C. Marr, Imbalanced domain generalization for robust single cell classification in hematological cytomorphology, 2023, CoRR. arXiv:2303.07771.

[56] S.W. Berson, HIPAA, Oncol. Issues 18 (2003) 20, URL https://api.semanticscholar.org/CorpusID:218541998.

[57] A.V. Parera, X. Costa, General data protection regulation, in: Data Protection Law in the EU: Roles, Responsibilities and Liability, 2018, URL https://api.semanticscholar.org/CorpusID:149781609.

[58] C. Liu, Y. Luo, Y. Xu, B. Du, HPFL: hyper-network guided personalized federated learning for multi-center tuberculosis chest x-ray diagnosis, Multimedia Tools Appl. (2023) 1–15.

[59] D.S. Sriram, A. Ranjan, V. Ghuge, N. Rathore, R. Agarwal, T. Diwan, J.V. Tembhurne, Personalized federated learning for the detection of COVID-19, Multimedia Tools Appl. (2023) 1–18.

[60] A. Sadilek, L. Liu, D. Nguyen, M. Kamruzzaman, S. Serghiou, B. Rader, A. Ingerman, S. Mellem, P. Kairouz, E.O. Nsoesie, J. Macfarlane, A. Vullikanti, M.V. Marathe, P. Eastham, J.S. Brownstein, B.A. y Arcas, M.D. Howell, J. Hernandez, Privacy-first health research with federated learning, Npj Digit. Med. 4 (2021).

[61] S. Sav, J. Bossuat, J.R. Troncoso-Pastoriza, M. Claassen, J. Hubaux, Privacy-preserving federated neural network learning for disease-associated cell classification, Patterns 3 (5) (2022) 100487.

[62] M.J. Sheller, B. Edwards, G.A. Reina, J. Martin, S. Pati, A. Kotrotsou, M. Milchenko, W. Xu, D. Marcus, R.R. Colen, et al., Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data, Sci. Rep. 10 (1) (2020) 12598.

[63] H.A. Madni, R.M. Umer, G.L. Foresti, Federated learning for data and model heterogeneity in medical imaging, in: G.L. Foresti, A. Fusiello, E. Hancock (Eds.), Image Analysis and Processing - ICIAP 2023 Workshops, Springer Nature Switzerland, Cham, 2024, pp. 167–178.

[64] M. Zhu, J. Liao, J. Liu, Y. Yuan, FedOSS: Federated open set recognition via inter-client discrepancy and collaboration, IEEE Trans. Med. Imaging 43 (1) (2024) 190–202.

[65] R.M. Umer, A. Gruber, S.S. Boushehri, C. Metak, C. Marr, Imbalanced domain generalization for robust single cell classification in hematological cytomorphology, 2023, CoRR. arXiv:2303.07771.

[66] A. Khosla, T. Zhou, T. Malisiewicz, A.A. Efros, A. Torralba, Undoing the damage of dataset bias, in: ECCV, 2012, pp. 158–171, http://dx.doi.org/10.1007/978-3-642-33718-5_12.

[67] C. Thapa, M.A.P. Chamikara, S.A. Camtepe, Advancements of federated learning towards privacy preservation: from federated learning to split learning, in: Federated Learning Systems: Towards Next-Generation AI, Springer, 2021, pp. 79–109.

[68] K. Gadzicki, R. Khamsehashari, C. Zetzsche, Early vs late fusion in multimodal convolutional neural networks, in: 2020 IEEE 23rd International Conference on Information Fusion, FUSION, 2020, pp. 1–6, http://dx.doi.org/10.23919/FUSION45008.2020.9190246.

[69] X. Huang, T. Ma, L. Jia, Y. Zhang, H. Rong, N. Alnabhan, An effective multimodal representation and fusion method for multimodal intent recognition, Neurocomputing 548 (2023) 126373, http://dx.doi.org/10.1016/j.neucom.2023.126373, URL https://www.sciencedirect.com/science/article/pii/S0925231223004964.

[70] J. Huertas-Tato, A. Martín, J. Fierrez, D. Camacho, Fusing CNNs and statistical indicators to improve image classification, Inf. Fusion 79 (2022) 174–187, http://dx.doi.org/10.1016/j.inffus.2021.09.012.

[71] R. Teixeira, L. Almeida, M. Antunes, D. Gomes, R.L. Aguiar, Efficient training: Federated learning cost analysis, Big Data Res. 40 (2025) 100510, http://dx.doi.org/10.1016/j.bdr.2025.100510, URL https://www.sciencedirect.com/science/article/pii/S221457962500005X.

[72] T. Li, A.K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, V. Smith, Federated optimization in heterogeneous networks, Proc. Mach. Learn. Syst. 2 (2020) 429–450.

[73] L. Su, J. Xu, P. Yang, A non-parametric view of FedAvg and FedProx: Beyond stationary points, 2021, arXiv preprint arXiv:2106.15216.

[74] B. Casella, R. Esposito, C. Cavazzoni, M. Aldinucci, Benchmarking FedAvg and FedCurv for image classification tasks, 2023, arXiv:2303.17942.

[75] H. Chen, J. Liu, C. Hua, J. Feng, B. Pang, D. Cao, C. Li, Accurate classification of white blood cells by coupling pre-trained ResNet and DenseNet with SCAM mechanism, BMC Bioinform. 23 (1) (2022) 282.

[76] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 770–778.

[77] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: F.R. Bach, D.M. Blei (Eds.), Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015, in: JMLR Workshop and Conference Proceedings, vol. 37, 2015, pp. 448–456, JMLR.org.

[78] A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, Q. Le, Searching for MobileNetV3, in: 2019 IEEE/CVF International Conference on Computer Vision, ICCV, 2019, pp. 1314–1324, http://dx.doi.org/10.1109/ICCV.2019.00140.

[79] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2015, pp. 1–9.

[80] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 2818–2826.

[81] G. Huang, Z. Liu, L. van der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, IEEE Computer Society, 2017, pp. 2261–2269.

[82] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, 2021, OpenReview.net.

[83] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: 2021 IEEE/CVF International Conference on Computer Vision, ICCV, 2021, pp. 9992–10002, http://dx.doi.org/10.1109/ICCV48922.2021.00986.

[84] S.H. Rezatofighi, H. Soltanian-Zadeh, Automatic recognition of five types of white blood cells in peripheral blood, Comput. Med. Imag. Graph. 35 (4) (2011) 333–343.

[85] M.R. Reena, P.M. Ameer, Segmentation of leukocyte by semantic segmentation model: A deep learning approach, Biomed. Signal Process. Control. 65 (2021) 102385.

[86] A. Acevedo, A. Merino, S. Alférez, Á. Molina, L. Boldú, J. Rodellar, A dataset of microscopic peripheral blood cell images for development of automatic recognition systems, Data Brief 30 (2020) 105474.

[87] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, Comput. Vis. Pattern Recognit. (CVPR) (2009) 248–255.

[88] H. Shin, H.R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D.J. Mollura, R.M. Summers, Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning, IEEE Trans. Med. Imaging 35 (5) (2016) 1285–1298.

[89] D. Beutel, et al., Flower: A friendly federated learning research framework, 2020, arXiv preprint arXiv:2007.14390.

[90] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (86) (2008) 2579–2605, URL http://jmlr.org/papers/v9/vandermaaten08a.html.

[91] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, IEEE Computer Society, 2017, pp. 618–626.

[92] M.B. Muhammad, M. Yeasin, Eigen-CAM: Class activation map using principal components, in: 2020 International Joint Conference on Neural Networks, IJCNN, 2020, pp. 1–7, http://dx.doi.org/10.1109/IJCNN48605.2020.9206626.

[93] V.J. Ramya, S. Lakshmi, Enhanced deep CNN based arithmetic optimization algorithm for acute myelogenous leukemia detection, Ann. Rom. Soc. Cell Biol. (2021) 7333–7352.

[94] T.A. Elhassan, M.S. Mohd Rahim, M.H. Siti Zaiton, T.T. Swee, T.A. Alhaj, A. Ali, M. Aljurf, Classification of atypical white blood cells in acute myeloid leukemia using a two-stage hybrid model based on deep convolutional autoencoder and deep convolutional neural network, Diagnostics 13 (2) (2023) 196.

[95] F. Ucar, Deep learning approach to cell classificatio in human peripheral blood, in: 2020 5th International Conference on Computer Science and Engineering, UBMK, IEEE, 2020, pp. 383–387.

[96] M. Abou Ali, F. Dornaika, I. Arganda-Carreras, Blood cell revolution: Unveiling 11 distinct types with 'naturalize'augmentation, Algorithms 16 (12) (2023) 562.

[97] Y.Y. Baydilli, Ü. Atila, Classification of white blood cells using capsule networks, Comput. Med. Imag. Graph. 80 (2020) 101699.

[98] P.P. Banik, R. Saha, K. Kim, An automatic nucleus segmentation and CNN model based classification method of white blood cell, Expert Syst. Appl. 149 (2020) 113211.

[99] E. Rivas-Posada, M.I.C. Murguia, Automatic base-model selection for white blood cell image classification using meta-learning, Comput. Biol. Med. 163 (2023) 107200.

[100] S. Dasariraju, M. Huo, S. McCalla, Detection and classification of immature leukocytes for diagnosis of acute myeloid leukemia using random forest algorithm, Bioengineering 7 (4) (2020) 120.

[101] T.A.M. Elhassan, M.S.M. Rahim, T.T. Swee, S.Z.M. Hashim, M. Aljurf, Feature extraction of white blood cells using CMYK-moment localization and deep learning in acute myeloid leukemia blood smear microscopic images, IEEE Access 10 (2022) 16577–16591.