Training and Benchmarking Leukocyte Sub-types Classification Methods with Synthetic Images

Luca Zedda[1[0009−0001−8488−1612]], Lorenzo Putzu[2[0000−0001−5361−8793]], Andrea Loddo[1[0000−0002−6571−3816]], and Cecilia Di Ruberto[1[0000−0003−4641−0307]]

[1] Department of Mathematics and Computer Science, University of Cagliari, Italy
[2] Department of Electrical and Electronic Engineering, University of Cagliari, Italy
{luca.zedda, lorenzo.putzu, andrea.loddo, cecilia.dir}@unica.it

Abstract. The classification of leukocyte sub-types is essential for medical diagnostics and treatments. Advances in this field have been driven by the creation of novel Deep Learning (DL) architectures, whose progress is sometimes marginal or not even comparable due to the use of proprietary data sets or different setups/partitions of public data sets. This study presents a novel synthetic image data set designed for both training and benchmarking, providing a standardised platform to evaluate advancements in this field. The data set includes two versions of differing complexity: straightforward and challenging. Experiments with various DL models showed unexpectedly higher accuracy, precision, and recall on the more complex data set. These results highlight the importance of data set complexity in assessing the robustness and effectiveness of DL models for complex medical image analysis tasks.

Keywords: Leukocyte classification · Image generation · Diffusion models

1 Introduction

Blood is composed of two main components: blood cells, which make up 45% of blood tissue volume, and plasma, accounting for the remaining 55% [8]. The three primary types of blood cells are platelets or thrombocytes, red blood cells (RBCs) or erythrocytes, and white blood cells (WBCs) or leukocytes. WBCs encompass five main sub-types: neutrophils, eosinophils, basophils, lymphocytes, and monocytes. Each plays a crucial role in the immune system by defending the body against microbial invaders and pathogens. Accurate classification of WBCs is vital for diagnosing hematopoietic diseases and guiding treatment decisions, particularly evident in conditions like leukaemia, where abnormal leukocyte proliferation occurs. Traditional methods of WBC classification, such as the manual microscopy conducted by trained experts, are essential for precise diagnosis but also time-consuming, error-prone, and subjective [64]. As a result, to avoid misdiagnosis, an automated procedure is becoming increasingly important, particularly in developing countries. The recent advancements in Machine Learning (ML), particularly Deep Learning (DL), have offered promising avenues for automated classification [20, 39, 61].

At present, DL algorithms have demonstrated potential in discerning and categorising leukocyte sub-types from a wide range of training data, necessitating a substantial number of accurate annotations acquired through labour-intensive efforts [7, 30, 39, 42, 51, 69]. However, publicly available real-world data sets are mostly limited in scale due to the expensive and hard manual labelling work, as well as privacy concerns. Moreover, real-world applications encounter challenges such as different lighting and acquisition conditions and variations in cell morphology and laboratory protocols, which existing data sets may not adequately cover [38]. These aspects may render conventional approaches ineffective at specific sites. Solutions like domain adaptation (DA) or domain generalisation (DG) have been proposed to enhance model robustness by leveraging data from the target domain or extracting domain-invariant features from multiple domains [46, 53, 65]. However, in a real-world application scenario, especially those related to medical images, due to privacy concerns, collecting and aggregating data from the target domain or more domains simultaneously is unfeasible [19, 58].

A potential solution to alleviate these issues is based on creating synthetic images, which can potentially be easily expanded to a large scale and address data scarcity, data imbalance, privacy concerns, and the high cost of manual data collection and annotation [45]. Synthetic data generation techniques can simulate various scenarios, augmenting existing data sets to encompass diverse conditions and variations, enriching the training data and enhancing ML models' robustness and generalisation capabilities. This enables researchers to explore novel domains where real-world data may be limited or inaccessible, offering a novel way to effectively pre-train or fine-tune Convolutional Neural Networks (CNNs) or Vision Transformers (ViTs), laying a stronger foundation for subsequent domains or applications [17, 67].

Inspired by the use of synthetic data sets for training DL models in different applications, we aim to propose a synthetic single-cell data set for WBC subtype classification. A further aim is to investigate if synthetic images could be reliably used for both training and benchmarking DL models. Indeed, current advancements in this field are mainly brought by the creation of novel architectures, whose progress in terms of performance is sometimes difficult to quantify due to the use of proprietary data sets (that cannot be shared for privacy concerns) or due to the use of different setups/partitions of public data sets, since most of them are small and do not present fixed training and testing partitions. Therefore, a shared standardised platform to evaluate advancements in this field is still lacking.

In order to perform such an investigation, we performed extensive same- and cross-data set experiments, including some existing real data sets and two synthetic data sets proposed in this work. These data sets have been created by exploiting Diffusion Models (DM), which are able to capture intricate patterns and provide a more precise representation of data distribution, particularly beneficial in biomedical inverse imaging, where they efficiently infer object or system properties while accommodating uncertainty and noise, promising advancements

in research and clinical applications [31]. We provided two data sets with two distinct complexity levels: one straightforward and the other more challenging. This endeavour aims to assess the influence on the construction of robust DL models and to evaluate their suitability for benchmarking purposes, effectively mitigating concerns related to real-world medical data. Moreover, the insights provided in this paper may offer inspiration for the design of future synthetic data sets, even across diverse applications.

The remaining sections of the manuscript are structured as follows. Sec. 2 present relevant previous research on WBC analysis, including the existing real benchmark data sets and generative models for medical images. Sec. 3 presents the proposed methodology and the creation of the proposed synthetic data set. Sec. 4 describes the DL architectures, the used real benchmark data sets and the experimental setting used in our evaluation, followed by the presentation of the experimental results and relative discussion. Lastly, Sec. 5 offers conclusions and considerations for future aspects.

2 Related Work

In this section, we first survey the literature on WBC sub-type classification (Section 2.1), followed by an analysis of available real data sets for such a task in Sec. 2.2. Finally, in Sec. 2.3, we present the existing approaches for image generation in the context of medical images.

2.1 White blood cell analysis

The diagnosis of hematopoietic malignancies, such as leukaemia, poses significant challenges due to the limited treatment options available. Traditionally, accurate diagnosis relies on meticulous cytological evaluation of WBCs in blood or bone marrow smears. However, the emergence of Computer-Aided Diagnosis (CAD) systems marks a transformative phase, offering automation and improved diagnostic accuracy and efficiency [3, 28, 38, 43, 62]. Existing CAD systems for WBC analysis cover a range of tasks, from simple cell counting to disease detection and classification. Notably, these systems excel in classifying WBC sub-types [1,20,43,48,51,59] and identifying specific leukaemia conditions, such as Acute Lymphoblastic Leukaemia (ALL) or Acute Myeloid Leukaemia (AML) [14, 38, 52].

For instance, Matek et al. utilised the ResNext architecture for accurately identifying WBCs in AML patients' blood smears [43]. Acevedo et al. proposed a predictive model for identifying patients with myelodysplastic syndrome, a precursor to AML [3]. Additionally, Vogado et al. leveraged transfer learning from multiple CNN architectures, coupled with SVM classifiers based on informative features derived from gain ratios, achieving remarkable results [62].

Overall, the primary advancements in the field stem from the creation of novel architectures, encompassing CNNs [20, 30, 39, 61, 69] or ViT [7, 42, 51], ensemble strategies [12], or hybrid methods [9,34,48,59], potentially incorporating feature

selection techniques [60]. A notable advancement in the field is exemplified by the development of BloodCaps by Long et al. , a capsule-based model tailored for the accurate classification of various blood cell types in peripheral blood images. BloodCaps outperformed traditional CNNs, such as AlexNet, VGG-16, ResNet-18, and Inception-V3, highlighting the potential of novel architectural paradigms to enhance classification techniques [39].

However, challenges persist, such as reliance on specific data sets, leading to performance degradation in case of a domain shift. In our previous work, we analysed the effectiveness of current WBC classification methods in real-world application scenarios, characterised by inaccurate regions of interest [38]. Our investigation unveiled that even simpler methodologies could yield accurate outcomes with suitable image training, as seen in benchmark data sets. However, only a handful of approaches, notably modern CNNs, demonstrated robustness in effectively tackling these challenges.

2.2 Existing Real Benchmark Data Sets

Table 1: Statistics of publicly available WBC data sets where #All is the total number of images while #5-WBC is the number of the 5 main WBC sub-types only.

Data set	#classes	**#L**	**#M**	**#N**	**#E**	**#B**	#5-WBC	#All imgs	Size (pixels)
LDWBC [13]	5	10,445	968	10,469	539	224	22,645	22,645	1280×1280
Raabin-WBC [32]	5	3,609	795	10,862	1,066	301	17,965	17,965	$\sim 575 \times 575$
LISC [49]	5	59	55	56	42	54	266	247	720×576
CellaVision [70]	5	37	18	30	12	3	100	100	300×300
JTSC [70]	5	53	48	176	22	1	300	300	120×120
Li et al. [35]	5	1,879	1,181	4,135	1,100	300	8,595	6,273	600×600
PBC [1, 2]	8	1,214	1,420	3,329	3,117	1,218	10,298	17,092	360×363
Bodzas et al. [11]	9	3,046	2,040	3,300	1,017	1,023	10,426	16,027	1200×1200
AML [43]	15	3,948	1,789	8,593	424	79	14,833	18,365	400×400

Table 1 presents a summary of the public WBC data sets proposed for WBC sub-types classification, indicating key insights like the number of classes, the composition and total of the primary five classes object of this study, the total number of images, and size in pixels. In certain instances, it is observed that the quantity of cells surpasses the number of available images. This discrepancy arises due to certain data sets (e.g., LISC and Li et al. [35]) lacking individual single-cell images. Instead, these data sets contain images where multiple cells may be present simultaneously.

Some public data sets proposed with different objectives were excluded from this table, such as BCCD [44], realised for distinguishing between WBCs, RBCs and platelets, or data sets such as ALL-IDB [33] and C-NMC [22] proposed to distinguish leukemic blast cells from healthy leukocytes.

Transitioning to the present data sets, a notable observation pertains primarily to a substantial variance in the number of represented classes. For instance, a significant portion of existing data sets consists of the five classes under study in

this research: lymphocytes (L), monocytes (M), neutrophils (N), eosinophils (E), and basophils (B), as evidenced in LDWBC, Raabin-WBC, LISC, CellaVision, and the data set proposed by Li et al. . Conversely, the remaining identified data sets encompass a greater number of classes, as PBC also represents immature granulocytes, erythroblasts, and platelets, while the data set by Bodzas et al. includes lymphoblasts, myeloblasts, normoblasts, and further classifies neutrophils into Band and Segmented. For simplicity, we have reported a single value derived from the sum of the latter two. Additionally, the AML data set comprises a total of 15 sub-types, with 8 representing immature cells and the remaining 7 representing mature cells.

Beyond the number of represented classes, a crucial aspect concerns the considerable imbalance among the compositions of various classes, which could significantly impact the performance of ML models trained on them. Specifically, disregarding data sets with few samples (LISC, CellaVision, and JTSC), it is evident that eosinophils and basophils, constituting a small proportion compared to other blood cells, are consequently underrepresented within the data sets. Furthermore, it is essential to note that the data sets were proposed without a predefined split into training, validation, and test sets. This lack of data division may impede an accurate assessment of model performance and generalisation on unseen data, as authors split the data arbitrarily, rendering final results incomparable. Despite these challenges, the data sets offer diverse characteristics, such as varying image dimensions and the total number of images. For example, the LDWBC data set comprises larger images at $1280x1280$ pixels.

In summary, while these public WBC data sets provide valuable resources for research and development in this context, their intrinsic class imbalances and the absence of predefined data splits underscore the importance of careful consideration and appropriate management when utilising them for ML tasks.

2.3 Generative Models for Medical Images

For the issues mentioned above, some authors have started to explore the generation of synthetic images in the context of WBC classification [10,23,36]. Generative models represent a transformative frontier within medical imaging, addressing critical challenges like data scarcity and inconsistency that often plague the field. Notably, Generative Adversarial Networks (GANs) have emerged, demonstrating remarkable efficacy across various domains within medical imaging [5]. In tasks ranging from data augmentation to registration and classification, GANs have showcased their adaptability and effectiveness, offering solutions to enhance image quality and facilitate analysis [16,41,63]. In [23], the fusion of autoencoders with a type of GAN, i.e., StyleGAN, was explored to generate four distinct types of leukocytes. Similarly, Liu et al. employs a different kind of GAN, i.e., Wasserstein GAN (WGAN), for the same objective [36]. Furthermore, Barrera et al. explored the use of two GAN architectures, namely WGAN and Super-Resolution Generative Adversarial Networks (SRGAN), for generating artificial images depicting both leukocytes and leukemic cells [10].

DM represents an alternative approach for image generation, circumventing the need for aligning posterior distributions, estimating intractable partition functions, introducing extra discriminator networks, or imposing network constraints [31]. DM has showcased state-of-the-art performance across diverse medical imaging applications, from segmentation to synthesis [4, 21]. One particularly promising avenue lies in the realm of anomaly detection, where DMs have exhibited notable success. By leveraging the inherent uncertainty modelling within diffusion processes, these models offer a principled approach to identifying anomalies in medical images, potentially revolutionising diagnostic processes and patient care [66].

3 Methodology

As mentioned before, the aim of this work is to create a new synthetic data set of single-cell images useful for the classification of the main five WBC sub-types. In the following, we provide a small background on the Denoising Diffusion Probabilistic Model (DDPM) used in this work (Sec. 3.1), preferred over other generative techniques for the reasons mentioned in Sec. 2.3. Then we describe the proposed methodology exploiting DDPM (Sec. 3.2) and the creation of synthetic images (Sec. 3.3).

3.1 DDPM

The DDPM is a generative model used for image-denoising tasks, where X denotes a clean image and Y a noisy image. The objective of DDPM is to model the conditional distribution $p(X|Y)$, which represents the probability of obtaining a clean image X given a noisy input Y. Formally, the objective is to train a model capable of progressively removing noise from noisy images over a series of time steps. Instead of directly mapping noisy images to their clean counterparts, the model is tasked with estimating the conditional distribution $p(X_t|Y_0)$, where X_t represents the clean image at timestep t and Y_0 is the initial noisy input. This distribution captures the evolving relationship between the clean and noisy images as noise is gradually added. The DDPM leverages a training data set containing pairs of clean and progressively noisier images to achieve this purpose. During training, the model's parameters are optimised to minimise the negative log-likelihood of this conditional distribution. By iteratively adjusting its predictions to match the clean images at each timestep, the DDPM effectively learns to generate realistic, denoised images from noisy inputs.

The U-Net architecture [50], originally proposed for image segmentation and widely adopted in various image processing tasks [6, 40] serves as a fundamental component even within the DDPM framework. It consists of an encoder-decoder structure with skip connections. The encoder portion extracts hierarchical features from the input image, while the decoder portion generates a high-resolution output by upsampling and combining features from different levels of abstraction. The skip connections facilitate effective information flow between corresponding

encoder and decoder layers, enabling the model to retain fine-grained details during upsampling.

3.2 Proposed Method

In order to create images that are realistic but at the same time have different levels of complexity so that they can also be used as a benchmark, we used an innovative procedure that consists of two main components: the DDPM and an image composition step. DDPMs are intended to generate new single cells. On the other hand, the image composition part is necessary to have more control over the number of cells in the final images and, thus, over the relative complexity of the image itself. A schema of the proposed methodology is shown in Fig. 1. In our implementation, we introduce an attention-enhanced U-Net

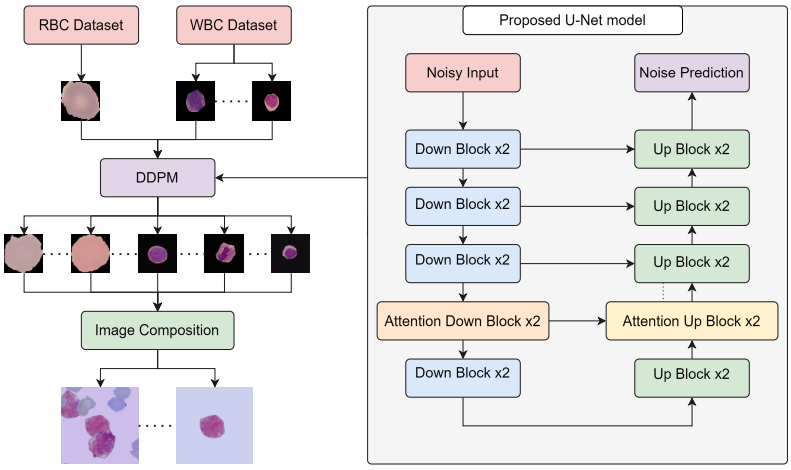

Fig. 1: Workflow of our pipeline for the synthetic data set creation and architecture of the custom attention-enhanced U-Net model. The custom U-Net model consists of 5 Downsample Blocks (Down Blocks) and 5 Upsample Blocks (Up Blocks), with each block repeated twice for enhanced feature extraction.

architecture, depicted in Fig. 1, which integrates Downsample Blocks and Upsample Blocks inspired by ResNet [25]. These blocks, repeated twice, are followed by convolutional downsampling and transposed convolutional upsampling operations to adjust the spatial resolution of the feature map. Attention blocks are strategically placed before downsampling and upsampling operations to capture intricate relationships within the feature maps. Each stage of the downsampling phase consists of blocks with 128, 128, 256, 256, and 512 channels, maintaining uniformity in the number of channels throughout the upsampling stage. While DDPM can be trained using a class-conditional configuration, where an embedding of the target class influences the generation process, there is a risk of the

model collapsing onto the majority class. Moreover, training a single model to generate images for both the training and testing partitions may result in synthetic samples closely resembling the same distribution, thereby compromising the validity of the test set outcomes. To mitigate these challenges, we partitioned the original source data sets into two subsets, each further segmented based on cell types (including 5 WBC sub-types and RBC), enabling the training of separate U-Net models for each class and data partition. We specifically chose the Raabin-WBC data set [32] as our source for WBC data and the Rajaraman RBC-only data set [47] for RBC data. Detailed information regarding the former data set is provided in Sec. 4.1, while the latter is publicly accessible and maintained by the National Institutes of Health (NIH). All images within the data set are manually annotated and encompass a total of 27,558 RBC images, evenly divided between malaria-parasitized and healthy RBCs [47]. For the sake of this work, we selected only healthy RBCs.

3.3 Creation of Synthetic Images

As mentioned, we produced two data sets with two very different complexity levels: Base Level (WBC-USID-B) and Complex Level (WBC-USID-C). The Base level depicts images with a single WBC cell against a static background. On the other hand, the Complex Level introduces greater complexity than the easy one, featuring other RBCs and WBCs in the background and occasionally providing partial occlusion with the central WBC. Our Complex Level also features a configurable level of background WBC and RBC displacement; in our work, we select our medium range of displacement 30% from the centre of the image with a standard deviation of 10%. Additionally, to prevent WBC and RBC alike from collapsing into single-point displacement, we reiterate the location decision until the next element does not have more than 85% overlap with other elements in the synthetic image. We selected a Base Level one as the starting point of a Complex Level image; this configuration aims to provide a more reliable comparison between the two sets by being a direct variety increment.

We produced a total of 2,000 samples with resolution 64×64 pixel for each complexity level. They both present training and testing partitions with 1,000 samples each, comprising 200 images for every WBC sub-type. All generated samples are publicly available on the official FigShare repository [68]. A sample image for each data set and each WBC sub-type is shown in Fig. 2.

4 Experimental Evaluation

In this section, we present the data sets (Sec. 4.1) and DL architectures employed in this study (Sec. 4.2); then, we outline the experimental setup (Sec. 4.3) and present and discuss the experimental results (Sec. 4.4) obtained.

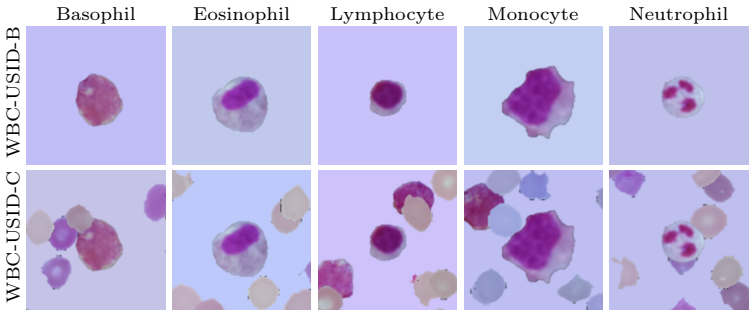

Fig. 2: A sample image of each WBC sub-type from the proposed synthetic data sets: Base (top row) and Complex (bottom row) levels row.

4.1 Data sets

In this study, we employed four established and publicly accessible data sets to have a rich set of benchmarks and comparison terms vs. our proposed synthetic data set. Given the limitations of existing data sets (mentioned in Sec. 2.2), we chose data sets covering a wider range of WBC classes with significant image quantities. We then pre-processed the data by removing or grouping classes to align with the specific classes relevant to our investigation.

The **Raabin-WBC** (from now on Raabin) data set comprises multiple subsets tailored to specific tasks. Our interest lies in a subset containing images with single WBCs, standardised to dimensions of 575×575 pixels. It comprises 1,145 images, showcasing 242 instances each of lymphocytes, monocytes, neutrophils, 201 eosinophils, and 218 basophils [32].

The **AML-Cytomorphology-LMU** (from now on AML) data set contains 18,365 images of dimensions 400×400 pixels, organised into 15 distinct categories. For this investigation, we focus on the five main WBC sub-types consisting of 1,789 instances of monocytes, 424 eosinophils, and 79 basophils. For the remaining two classes, we merged 8,484 segmented neutrophils and 109 band neutrophils into a single neutrophil category, 3,937 typical lymphocytes and 11 atypical lymphocytes into a lymphocyte category [43].

The **PBC** data set comprises 17,092 images of size 360×363 pixels subdivided into eight categories: neutrophils, eosinophils, basophils, lymphocytes, monocytes, immature granulocytes, erythroblasts, and peripheral blood cell types. In this study, we selected five classes, with the following counts: 3,329 instances of neutrophils, 3,117 of eosinophils, 1,218 of basophils, 1,214 of lymphocytes, and 1,420 of monocytes [1, 2].

The **LDWBC** data set comprises 22,645 images, categorised as basophil (224), monocyte (968), eosinophil (539), neutrophil (10,469), and lymphocyte (10,445). Each image has a resolution of 1280×1280 pixels [13].

A sample image for each data set and each WBC sub-type is shown in Sec. 4.1.

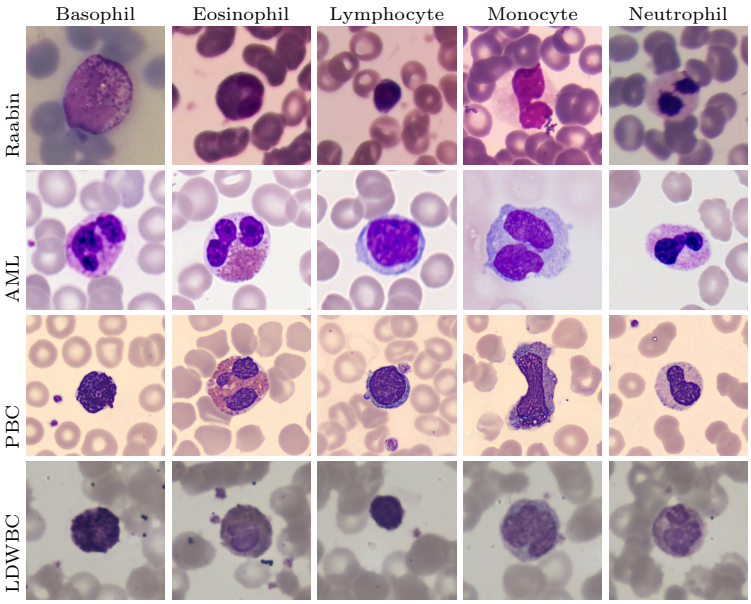

Fig. 3: A sample image of each WBC sub-type from the four real data sets utilised in this study: Raabin, AML, PBC, and LDWBC.

4.2 Deep Learning architectures

Our primary aim in this experimentation is to evaluate and compare different DL architectures for the classification of WBC sub-types in two different scenarios: real-world data sets and our proposed synthetic WBC data set. To ensure the reliability of our findings despite variations in network architecture, such as differences in layer counts, the presence of skip connections, and other architectural nuances, we conducted evaluations across a range of DL architectures. Given the extensive variety of CNNs proposed to date, alongside recent advancements in ViTs tailored for classification tasks, we selected one representative architecture from each of the following categories: linear, residual, lightweight, densely connected and inception CNNs, vision transformers, and hierarchical vision transformers. From this selection, the chosen architectures include VGG-19, ResNet-152, MobileNetV3, DenseNet-121, Inception-V3, ViT, and Swin.

VGG-19 architecture is a linear network that comprises 19 layers and has been recognised for its simplicity and directness [55].

ResNet-152 has 152 layers and incorporates the residual learning principles. These principles involve the integration of skip connections or recurrent units between convolution and pooling layers, alongside batch normalisation for enhanced training stability [24, 29].

MobileNet-V3 belongs to the MobileNet family, designed for mobile and embedded systems. This family's models utilise depthwise separable convolutions

to replace conventional convolution layers, reducing computational complexity and model size while maintaining accuracy [26].

Inception-V3, an evolution of GoogLeNet, is built upon inception layers comprising various convolutional filters. This design facilitates multi-scale feature learning and employs global average pooling, factorised convolutions, and auxiliary classifiers for regularisation [56, 57].

DenseNet addresses limitations in traditional CNNs by introducing dense connections between layers, where each layer receives input from all preceding layers. The number of filters in each convolution layer dynamically adjusts based on the growth rate parameter, influencing the overall parameter count [27]. Here, we used **DenseNet-121**.

ViT represents a transformer-based architecture for computer vision tasks, leveraging self-attention mechanisms instead of conventional convolutions. Images are divided into fixed-size patches and processed sequentially through transformer layers, allowing the model to capture spatial information effectively [18].

Swin, an advancement over ViT, addresses long-range dependency challenges by employing hierarchical self-attention. It utilises a shifted window approach for patch division, enabling overlapping patches to capture richer contextual information. Features are aggregated using Swin blocks, encompassing both local and global dependencies [37].

4.3 Experimental Setup

The DL architectures used in this study were initially pre-trained on the natural image data set, ImageNet [15], before being adapted to medical image tasks. This adaptation process followed a well-established procedure for transfer learning and fine-tuning CNN models, as outlined in prior literature [54]. All CNN layers were preserved during this adaptation 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.

Two different experiments have been performed. In the first one, a classic same-data set setup is used, where both real and synthetic data sets are used for both training and testing the DL models. This experiment is meant to demonstrate if the proposed synthetic data set can be used for benchmarking by assessing if the obtained results are similar among all the used DL architectures. In the second experiment, instead, both real and synthetic data sets are used in a cross-data set fashion, meaning that each data set is alternately designated as the SOURCE (training) data set, and the obtained model is evaluated an all TARGET (testing) data sets. This experiment is meant to assess the generalisation capabilities of the proposed synthetic data sets compared with the generalisation capabilities of existing real data sets.

Given that the used real data sets do not present training and testing partitions, to ensure a fair comparison and enable direct comparison on the same testing partition, all mentioned real data sets were partitioned into three fixed subsets: training, validation, and testing sets, constituting 60%, 20%, and 20% of the original data set sizes, respectively. For the proposed synthetic data sets

that already present training and testing partitions, we split the original training partition into training and validation subsets constituting 75% and 25% of the original partition, respectively. A stratified sampling procedure was employed for both real and synthetic data sets to maintain the original data distribution. Given the inherent class imbalance in some real data sets, a weighted random sampling technique was incorporated during training to ensure equal representation of each class within each batch. Furthermore, data augmentation procedures, including random rotation, random horizontal and vertical flipping, centre cropping, Gaussian blurring, and colour jitter, were applied to mitigate the risk of overfitting during the training phase.

All experiments were conducted on a single machine featuring an Intel(R) Xeon(R) CPU E5-2670 v3 @ 2.30GHz CPU, 128 GB RAM, and NVIDIA Quadro 24GB GPU. Training processes were executed for a maximum of 100 epochs, with early stopping based on validation performance. A batch size of 32, a Stochastic Gradient Descent (SGD) optimiser, and a learning rate of 0.001 were employed throughout the training procedures. Performance was measured with Accuracy (A), Precision (P), Recall (R), and F1-score (F1). These metrics are calculated individually for each class and subsequently aggregated using a weighted average procedure to derive a consolidated performance measure.

4.4 Experimental Results

Table 2 presents the same-data set results obtained on both real and synthetic data sets using the considered DL architectures. We analyse and compare their results to those on real data sets to understand the reliability of benchmarking on the proposed synthetic data sets. As expected, the results are very similar, with some differences that can be attributed to data set bias and other influential factors. Indeed, even when comparing pairs of real data sets, it is not possible to determine a single DL architecture that outperforms the others. However, when comparing the two synthetic data sets, where the bias and influential factors are limited by design (see Sec. 3.3) a very similar trend is observed, underscoring the importance of synthetic data sets even for benchmarking, given that they allow extracting a leaderboard of the best DL model on the addressed task. A further observation emerges from the examination of the synthetic data sets: in numerous instances, the performance attained with the Complex Level surpasses that of its simpler counterpart. This phenomenon might be attributed to the presence of additional objects (cells) alongside the primary object of interest, compelling the models to concentrate more on it and its intricate characteristics. Table 3 present the cross-data set results. Here, we also include same-data set performance as a baseline for comparison, indicated with a grey background in the tables. This allows for a direct assessment of the deviation in performance between same-data set and cross-data set scenarios. From this assessment, it is plausible to expect a much larger decrease in real TARGET data sets when synthetic data sets are used as SOURCE, as they have very different visual characteristics. Nevertheless, in many cases, the performance is aligned with the cases when the real data sets are used as SOURCE and, in some cases, even higher; see Raabin TARGET

	Raabin				AML				PBC				LDWBC				WBC-USID-B				WBC-USID-C			
Models	A	P	R	F1	A	P	R	F1	A	P	R	F1	A	P	R	F1	A	P	R	F1	A	P	R	F1
VGG-19	96.5	96.6	96.5	96.5	88.1	89.4	88.1	87.9	97.1	97.3	97.1	97.1	89.8	92.4	89.8	89.8	96.1	96.1	96.1	96.1	96.9	97.2	96.9	96.9
ResNet-152	94.3	94.7	94.3	94.3	87.1	89.4	87.1	86.7	**99.2**	**99.2**	**99.2**	**99.2**	95.1	95.8	95.1	95.2	96.5	96.6	96.5	96.5	**98.4**	**98.4**	**98.4**	**98.4**
Inception-V3	96.9	97.0	96.9	96.9	86.6	89.6	86.6	86.0	95.0	96.0	95.0	95.0	94.2	94.8	94.2	94.3	**99.1**	**99.1**	**99.1**	**99.1**	**98.4**	**98.4**	**98.4**	**98.4**
DenseNet-121	85.6	89.4	85.6	85.0	68.6	78.5	68.6	65.5	97.9	98.1	97.9	98.0	96.4	96.7	96.4	96.4	95.7	96.0	95.7	95.7	94.8	95.3	94.8	94.7
MobileNet-V3	96.1	96.3	96.1	96.1	91.8	91.7	91.8	91.7	95.5	95.9	95.5	95.5	93.3	94.2	93.3	93.4	89.7	90.4	89.7	89.6	95.3	95.3	95.3	95.3
ViT	94.8	94.9	94.8	94.7	**93.3**	**93.5**	**93.3**	**93.3**	96.7	96.8	96.7	96.7	**97.3**	**97.5**	**97.3**	**97.3**	93.2	93.5	93.2	93.2	94.3	94.5	94.3	94.3
Swin	**97.8**	**97.8**	**97.8**	**97.8**	88.7	90.6	88.7	88.3	97.9	98.0	97.9	97.9	93.3	94.5	93.3	93.2	95.9	96.1	95.9	95.9	95.9	96.1	95.9	95.9

Table 2: Same-data results on real and synthetic data sets. The best and second-best performances for each data set are indicated in bold and underlined, respectively.

		TARGET																							
Models		Raabin				AML				PBC				LDWBC				WBC-USID-B				WBC-USID-C			
	SOURCE	A	P	R	F1	A	P	R	F1	A	P	R	F1	A	P	R	F1	A	P	R	F1	A	P	R	F1
VGG-19	Raabin	96.5	96.6	96.5	96.5	63.4	70.7	63.4	64.6	57.9	55.7	57.9	53.1	58.2	65.3	58.2	59.7	77.6	79.4	77.6	77.1	66.3	70.8	66.3	64.7
	AML	54.1	62.8	54.1	48.4	88.1	89.4	88.1	87.9	42.6	57.5	42.6	38.1	42.7	39.7	42.7	38.5	35.3	58.1	35.3	25.5	25.4	34.4	25.4	16.4
	PBC	65.5	72.0	65.5	58.9	83.5	84.0	83.5	83.6	97.1	97.3	97.1	97.1	67.6	68.7	67.6	64.6	62.1	55.0	62.1	56.1	49.8	43.8	49.8	43.4
	LDWBC	51.5	55.5	51.5	45.6	78.9	80.3	78.9	77.4	61.2	68.1	61.2	56.9	89.8	92.4	89.8	89.8	45.7	68.9	45.7	41.0	33.4	51.3	33.4	30.1
	WBC-USID-B	48.9	74.8	48.9	50.2	44.8	66.7	44.8	43.6	31.4	53.3	31.4	26.4	36.4	56.4	36.4	33.7	96.1	96.1	96.1	96.1	76.6	88.2	76.6	76.1
	WBC-USID-C	50.7	60.9	50.7	48.2	32.0	61.6	32.0	34.9	25.6	38.5	25.6	15.4	26.7	36.4	26.7	19.1	98.1	98.2	98.1	98.1	96.9	97.2	96.9	96.9
ResNet-152	Raabin	94.3	94.7	94.3	94.3	76.8	77.7	76.8	77.1	40.1	50.6	40.1	32.4	53.3	55.5	53.3	49.0	57.7	76.4	57.7	60.2	53.2	63.9	53.2	53.0
	AML	64.2	79.9	64.2	56.4	87.1	89.4	87.1	86.7	43.8	70.3	43.8	34.5	50.7	52.3	50.7	45.8	31.1	35.9	31.1	19.9	23.9	46.5	23.9	13.5
	PBC	63.3	67.8	63.3	60.2	73.2	80.5	73.2	72.9	99.2	99.2	99.2	99.2	79.1	81.1	79.1	78.2	57.7	54.1	57.7	52.1	45.6	45.1	45.6	40.6
	LDWBC	44.5	43.5	44.5	36.0	66.5	78.7	66.5	62.8	52.5	54.0	52.5	44.2	95.1	95.8	95.1	95.2	32.8	39.4	32.8	20.9	21.3	12.8	21.3	9.4
	WBC-USID-B	38.0	70.4	38.0	37.9	46.4	71.3	46.4	46.3	53.7	66.1	53.7	49.6	40.9	41.2	40.9	38.4	96.5	96.6	96.5	96.5	53.0	79.7	53.0	53.1
	WBC-USID-C	69.0	79.5	69.0	65.7	46.9	67.1	46.9	51.9	33.1	38.6	33.1	26.4	33.3	54.8	33.3	29.7	91.4	92.9	91.4	91.1	98.4	98.4	98.4	98.4
Inception-V3	Raabin	96.9	97.0	96.9	96.9	71.1	77.7	71.1	73.3	47.5	47.5	47.5	39.5	51.6	60.4	51.6	52.1	37.0	64.4	37.0	33.5	54.8	59.9	54.8	54.8
	AML	62.9	73.7	62.9	59.3	86.6	89.6	86.6	86.0	40.9	68.2	40.9	32.3	45.8	50.7	45.8	41.8	28.7	69.6	28.7	18.9	24.4	40.1	24.4	14.8
	PBC	65.5	74.5	65.5	62.8	77.8	81.3	77.8	76.7	95.0	96.0	95.0	95.0	59.1	61.7	59.1	59.0	54.9	53.5	54.9	49.3	41.7	42.6	41.7	33.0
	LDWBC	45.0	51.0	45.0	37.6	67.5	75.6	67.5	65.5	25.2	15.8	25.2	14.1	94.2	94.8	94.2	94.3	38.0	50.1	38.0	29.9	24.8	35.8	24.8	16.6
	WBC-USID-B	24.9	34.2	24.9	17.6	27.3	53.5	27.3	27.8	28.5	27.3	28.5	21.4	30.2	46.9	30.2	21.1	99.1	99.1	99.1	99.1	86.4	88.6	86.4	86.5
	WBC-USID-C	45.4	46.9	45.4	40.6	30.9	64.4	30.9	35.7	29.8	15.0	29.8	17.8	23.1	37.6	23.1	12.7	97.2	97.3	97.2	97.2	98.4	98.4	98.4	98.4
DenseNet-121	Raabin	85.6	89.4	85.6	85.0	77.8	77.7	77.8	76.0	41.3	55.6	41.3	34.2	45.3	62.6	45.3	41.0	56.5	63.2	56.5	55.7	42.2	57.8	42.2	42.0
	AML	40.2	60.4	40.2	31.9	68.6	78.5	68.6	65.5	27.3	11.0	27.3	15.5	40.9	38.2	40.9	33.0	39.6	66.6	39.6	29.5	21.4	38.6	21.4	9.6
	PBC	63.8	67.6	63.8	60.5	74.7	80.7	74.7	74.4	97.9	98.1	97.9	97.9	61.3	62.2	61.3	59.4	56.8	55.5	56.8	49.7	37.3	46.8	37.3	33.4
	LDWBC	51.5	62.0	51.5	45.3	78.4	82.4	78.4	75.8	82.2	85.7	82.2	81.2	96.4	96.7	96.4	96.4	25.0	32.8	25.0	16.1	20.1	16.8	20.1	7.7
	WBC-USID-B	30.6	50.2	30.6	25.5	29.9	68.8	29.9	30.4	41.3	42.6	41.3	35.3	35.6	58.2	35.6	31.5	95.7	96.0	95.7	95.7	64.5	77.0	64.5	63.1
	WBC-USID-C	48.0	54.4	48.0	44.9	45.4	75.7	45.4	50.3	38.8	61.7	38.8	28.2	36.0	60.6	36.0	33.7	98.3	98.4	98.3	98.3	94.8	95.3	94.8	94.7
MobileNet-V3	Raabin	96.1	96.3	96.1	96.1	44.3	65.6	44.3	43.9	55.0	55.6	55.0	48.4	47.6	62.1	47.6	46.3	39.9	56.9	39.9	37.3	58.3	65.4	58.3	58.7
	AML	53.7	58.4	53.7	50.9	91.8	91.7	91.8	91.7	71.1	78.7	71.1	70.1	39.1	50.6	39.1	33.0	47.1	53.8	47.1	40.2	29.4	50.9	29.4	23.5
	PBC	47.6	38.9	47.6	38.7	69.1	73.9	69.1	66.6	95.5	95.9	95.5	95.5	57.8	71.9	57.8	52.8	33.9	49.2	33.9	28.2	25.3	35.3	25.3	17.7
	LDWBC	32.3	47.7	32.3	31.8	52.1	47.5	52.1	46.3	33.5	47.8	33.5	25.4	93.3	94.2	93.3	93.4	25.7	33.1	25.7	21.5	26.6	41.1	26.6	19.8
	WBC-USID-B	38.0	44.1	38.0	36.6	35.1	51.6	35.1	34.0	16.5	27.2	16.5	9.5	30.2	30.8	30.2	24.5	89.7	90.4	89.7	89.6	77.9	79.1	77.9	78.1
	WBC-USID-C	47.6	47.3	47.6	42.6	27.3	39.5	27.3	27.9	27.3	39.2	27.3	19.9	21.8	30.2	21.8	13.3	91.6	93.1	91.6	91.3	95.3	95.3	95.3	95.3
ViT	Raabin	94.8	94.9	94.8	94.7	56.7	67.1	56.7	56.7	59.9	58.4	59.9	54.5	55.1	66.3	55.1	53.3	60.9	65.0	60.9	62.3	46.8	55.6	46.8	41.8
	AML	65.5	67.0	65.5	61.6	93.3	93.5	93.3	93.3	79.3	83.4	79.3	78.9	62.7	75.1	62.7	61.9	44.7	52.2	44.7	40.8	40.9	63.2	40.9	34.7
	PBC	68.1	74.0	68.1	63.9	64.9	79.0	64.9	65.2	96.7	96.8	96.7	96.7	66.2	69.9	66.2	65.2	48.2	50.4	48.2	39.7	40.0	48.1	40.0	37.1
	LDWBC	59.8	63.6	59.8	57.9	58.8	70.0	58.8	56.0	63.6	77.7	63.6	61.7	97.3	97.5	97.3	97.3	43.4	42.5	43.4	39.0	31.5	44.4	31.5	24.2
	WBC-USID-B	38.9	75.6	38.9	36.3	37.1	63.2	37.1	31.6	51.2	65.0	51.2	49.4	40.9	52.6	40.9	36.6	93.2	93.5	93.2	93.2	75.0	83.1	75.0	76.0
	WBC-USID-C	58.5	78.1	58.5	56.4	32.0	66.9	32.0	35.4	41.7	61.7	41.7	36.6	29.8	54.2	29.8	24.1	91.8	92.3	91.8	91.7	94.3	94.5	94.3	94.3
Swin	Raabin	97.8	97.8	97.8	97.8	62.9	70.2	62.9	63.3	60.3	61.8	60.3	54.2	52.9	56.5	52.9	50.2	77.6	81.4	77.6	78.0	63.2	70.1	63.2	60.3
	AML	63.3	66.4	63.3	55.6	88.7	90.6	88.7	88.3	52.9	72.9	52.9	43.4	44.9	40.5	44.9	35.8	26.6	33.1	26.6	17.9	24.4	52.6	24.4	15.9
	PBC	67.7	77.1	67.7	61.5	69.6	75.8	69.6	67.5	97.9	98.0	97.9	97.9	60.4	71.2	60.4	57.0	47.1	57.4	47.1	42.7	36.6	39.3	36.6	33.6
	LDWBC	71.2	74.2	71.2	69.6	58.8	71.9	58.8	51.0	71.9	82.6	71.9	71.4	93.3	94.5	93.3	93.2	65.5	64.9	65.5	61.2	32.5	45.4	32.5	23.4
	WBC-USID-B	60.7	81.1	60.7	57.2	54.6	72.1	54.6	55.2	54.5	68.3	54.5	47.8	43.6	53.2	43.6	39.0	95.9	96.1	95.9	95.9	76.2	87.7	76.2	77.2
	WBC-USID-C	58.5	79.4	58.5	53.3	42.3	73.8	42.3	47.2	51.2	53.8	51.2	45.8	33.3	42.9	33.3	28.9	99.0	99.0	99.0	99.0	95.9	96.1	95.9	95.9

Table 3: Same-data (highlighted in grey) and cross-data set performance on TARGET data sets exploiting the model trained with the SOURCE data set.

data sets with ResNet-152. This is the most interesting result, considering no additional fine-tuning was performed with real images. The most commonly used approach in the literature, indeed, is to fine-tune the model again after training with synthetic images or directly mixing real and synthetic images because it is more intuitive to guide real-world applications.

Further observations emerge from examining the columns related to synthetic data sets. In such scenarios, a distinct decline in performance becomes apparent solely when utilising the Base Level as the SOURCE and the Complex one as the TARGET. Conversely, no decrease in performance is observed when the Complex Level serves as the SOURCE and the Base one as the TARGET; in fact, in numerous cross-data set evaluations, performance actually improves. This further underscores how escalating complexity in the images facilitates the training of more resilient models with enhanced generalisation capabilities. These findings demonstrate that the proposed data sets are suitable for direct training of DL models, ensuring that in real-world applications, they are characterised by very high bias and perform similarly to models trained with real images. In addition, the proposed synthetic data set generation method presents numerous significant advantages over the literature. These range from the privacy-preserving aspect of utilising synthetic images to the substantial time saved on annotations with respect to the use of GANs [10, 23, 36]. This efficiency is attributed to the capability of the proposed methodology, based on DDPM, to generate both images and annotations simultaneously in a single step and the ability to handle image complexity while minimising bias.

5 Conclusion

In conclusion, this study presents a novel synthetic image data set specifically designed to train and evaluate DL methods for classifying leukocyte sub-types critical for various medical applications. The data set was generated using the DDPM, which ensures high-quality synthetic images. Two meticulously crafted versions of the data set were developed, each varying in complexity. Experimental results across several DL architectures demonstrate a significant difference in model performance between the two versions, with notably higher accuracy, precision, and recall achieved on the Complex Level data set. These findings underscore the importance of data set complexity in assessing the robustness and effectiveness of DL models for complex medical image analysis tasks and confirm that the proposed data set can serve as a valuable resource for researchers, offering a standardised platform to benchmark and drive advancements in WBC sub-type classification, thereby advancing medical research and healthcare applications. Additionally, the study in this paper might also inspire future designs of synthetic data sets and advance further studies with synthetic data sets for even wider applications in the future. While this study acknowledges limitations in the quantity and resolution of produced images, and the analysis is performed on two complexity levels only, future research endeavours aim to address these issues by producing a large-scale data version with higher resolution and by introducing intermediate complexity levels gradually incorporating different types of cells. This approach is expected to provide a more nuanced understanding of the correlation between data set size, quality, complexity and model performance, further enhancing the efficacy of DL methods in medical image analysis.

References

1. Acevedo, A., Alférez, S., Merino, A., Puigví, L., Rodellar, J.: Recognition of peripheral blood cell images using convolutional neural networks. Computer Methods and Programs in Biomedicine **180**, 105020 (2019)
2. Acevedo, A., Merino, A., Alférez, S., Ángel Molina, Boldú, L., Rodellar, J.: A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. Data in Brief **30**, 105474 (2020)
3. Acevedo, A., Merino, A., Boldú, L., Molina, A., Alférez, S., Rodellar, J.: A new convolutional neural network predictive model for the automatic recognition of hypogranulated neutrophils in myelodysplastic syndromes. Comput. Biol. Medicine **134**, 104479 (2021)
4. Akrout, M., Gyepesi, B., Holló, P., Poór, A., Kincso, B., Solis, S., Cirone, K., Kawahara, J., Slade, D., Abid, L., Kovács, M., Fazekas, I.: Diffusion-based data augmentation for skin disease classification: Impact across original medical datasets to fully synthetic images. In: Deep Generative Models. pp. 99–109. Springer Nature Switzerland, Cham (2024)
5. AlAmir, M., Ghamdi, M.A.: The role of generative adversarial network in medical image analysis: An in-depth survey. ACM Comput. Surv. **55**(5), 96:1–96:36 (2023)
6. Alharbi, A.H., Aravinda, C.V., Lin, M., Venugopala, P.S., Reddicherla, P., Shah, M.A.: Segmentation and Classification of White Blood Cells Using the UNet. Contrast Media & Molecular Imaging **2022**, 5913905 (Jul 2022). `https://doi.org/10.1155/2022/5913905`, `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9293541/`
7. Almalik, F., Alkhunaizi, N., Almakky, I., Nandakumar, K.: Fesvibs: Federated split learning of vision transformer with block sampling. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T.F., Taylor, R.H. (eds.) Medical Image Computing and Computer Assisted Intervention - MICCAI 2023 - 26th International Conference, Vancouver, BC, Canada, October 8-12, 2023, Proceedings, Part II. Lecture Notes in Computer Science, vol. 14221, pp. 350–360. Springer (2023)
8. Bain, B.J.: Blood cells: a practical guide. John Wiley & Sons (2021)
9. Bairaboina, S.S.R., Battula, S.R.: Ghost-resnext: An effective deep learning based on mature and immature wbc classification. Applied Sciences **13**(6), 4054 (2023)
10. Barrera, K., Merino, A., Molina, A., Rodellar, J.: Automatic generation of artificial images of leukocytes and leukemic cells using generative adversarial networks (syntheticcellgan). Comput. Methods Programs Biomed. **229**, 107314 (2023)
11. Bodzas, A., Kodytek, P., Zidek, J.: A high-resolution large-scale dataset of pathological and normal white blood cells. Scientific Data **10**(1), 466 (2023)
12. Bravin, R., Nanni, L., Loreggia, A., Brahnam, S., Paci, M.: Varied image data augmentation methods for building ensemble. IEEE Access **11**, 8810–8823 (2023)
13. Chen, H., Liu, J., Hua, C., Feng, J., Pang, B., Cao, D., Li, C.: Accurate classification of white blood cells by coupling pre-trained resnet and densenet with SCAM mechanism. BMC Bioinform. **23**(1), 282 (2022)
14. Das, P.K., Meher, S.: An efficient deep convolutional neural network based detection and classification of acute lymphoblastic leukemia. Expert Syst. Appl. **183**, 115311 (2021)
15. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. Computer Vision and Pattern Recognition (CVPR) pp. 248–255 (2009)

16. Desai, S.D., Giraddi, S., Verma, N., Gupta, P., Ramya, S.: Breast cancer detection using gan for limited labeled dataset. In: 2020 12th International Conference on Computational Intelligence and Communication Networks (CICN). pp. 34–39. IEEE (2020)
17. Ding, K., Zhou, M., Wang, H., Gevaert, O., Metaxas, D., Zhang, S.: A large-scale synthetic pathological dataset for deep learning-enabled segmentation of breast cancer. Scientific Data **10**(1), 231 (2023)
18. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: 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. OpenReview.net (2021)
19. Felfeliyan, B., Forkert, N.D., Hareendranathan, A.R., Cornel, D., Zhou, Y., Kuntze, G., Jaremko, J.L., Ronsky, J.L.: Self-supervised-rcnn for medical image segmentation with limited data annotation. Comput. Medical Imaging Graph. **109**, 102297 (2023)
20. Fırat, H.: Classification of microscopic peripheral blood cell images using multi-branch lightweight cnn-based model. Neural Computing and Applications **36**(4), 1599–1620 (2024)
21. Güngör, A., Dar, S.U.H., Öztürk, S., Korkmaz, Y., Bedel, H.A., Elmas, G., Özbey, M., Çukur, T.: Adaptive diffusion priors for accelerated MRI reconstruction. Medical Image Anal. **88**, 102872 (2023). https://doi.org/10.1016/J.MEDIA.2023.102872, https://doi.org/10.1016/j.media.2023.102872
22. Gupta, R., Gehlot, S., Gupta, A.: C-nmc: B-lineage acute lymphoblastic leukaemia: A blood cancer dataset. Medical Engineering & Physics **103**, 103793 (2022)
23. Hazra, D., Byun, Y.C., Kim, W.J., Kang, C.U.: Synthesis of microscopic cell images obtained from bone marrow aspirate smears through generative adversarial networks. Biology **11**(2), 276 (2022)
24. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
25. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778. IEEE, Las Vegas, NV, USA (Jun 2016). https://doi.org/10.1109/CVPR.2016.90, http://ieeexplore.ieee.org/document/7780459/
26. Howard, A., Sandler, M., Chen, B., Wang, W., Chen, L.C., Tan, M., Chu, G., Vasudevan, V., Zhu, Y., Pang, R., Adam, H., Le, Q.: Searching for mobilenetv3. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1314–1324 (2019). https://doi.org/10.1109/ICCV.2019.00140
27. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 2261–2269. IEEE Computer Society (2017)
28. Huang, Q., Li, W., Zhang, B., Li, Q., Tao, R., Lovell, N.H.: Blood cell classification based on hyperspectral imaging with modulated gabor and cnn. IEEE Journal of Biomedical and Health Informatics **24**(1), 160–170 (jan 2020)
29. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Bach, F.R., Blei, D.M. (eds.) Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015. JMLR Workshop and Conference Proceedings, vol. 37, pp. 448–456. JMLR.org (2015)

30. Jiang, L., Tang, C., Zhou, H.: White blood cell classification via a discriminative region detection assisted feature aggregation network. Biomedical Optics Express **13**(10), 5246–5260 (2022)
31. Kazerouni, A., Aghdam, E.K., Heidari, M., Azad, R., Fayyaz, M., Hacihaliloglu, I., Merhof, D.: Diffusion models in medical imaging: A comprehensive survey. Medical Image Anal. **88**, 102846 (2023). https://doi.org/10.1016/J.MEDIA.2023.102846, https://doi.org/10.1016/j.media.2023.102846
32. Kouzehkanan, Z.M., Saghari, S., Tavakoli, S., Rostami, P., Abaszadeh, M., Mirzadeh, F., Satlsar, E.S., Gheidishahran, M., Gorgi, F., Mohammadi, S., et al.: A large dataset of white blood cells containing cell locations and types, along with segmented nuclei and cytoplasm. Scientific reports **12**(1), 1123 (2022)
33. Labati, R.D., Piuri, V., Scotti, F.: All-idb: The acute lymphoblastic leukemia image database for image processing. In: Macq, B., Schelkens, P. (eds.) 18th IEEE International Conference on Image Processing, ICIP 2011, Brussels, Belgium, September 11-14, 2011. pp. 2045–2048. IEEE (2011)
34. Li, C., Liu, Y.: Improved generalization of white blood cell classification by learnable illumination intensity invariant layer. IEEE Signal Process. Lett. **31**, 176–180 (2024)
35. Li, M., Lin, C., Ge, P., Li, L., Song, S., Zhang, H., Lu, L., Liu, X., Zheng, F., Zhang, S., et al.: A deep learning model for detection of leukocytes under various interference factors. Scientific Reports **13**(1), 2160 (2023)
36. Liu, K., Shuai, R., Ma, L., et al.: Cells image generation method based on vae-sgan. Procedia Computer Science **183**, 589–595 (2021)
37. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9992–10002 (2021). https://doi.org/10.1109/ICCV48922.2021.00986
38. Loddo, A., Putzu, L.: On the effectiveness of leukocytes classification methods in a real application scenario. AI **2**(3), 394–412 (2021). https://doi.org/10.3390/ai2030025, https://www.mdpi.com/2673-2688/2/3/25
39. Long, F., Peng, J., Song, W., Xia, X., Sang, J.: Bloodcaps: A capsule network based model for the multiclassification of human peripheral blood cells. Comput. Methods Programs Biomed. **202**, 105972 (2021)
40. Lu, Y., Qin, X., Fan, H., Lai, T., Li, Z.: WBC-Net: A white blood cell segmentation network based on UNet++ and ResNet. Applied Soft Computing **101**, 107006 (Mar 2021). https://doi.org/10.1016/j.asoc.2020.107006, https://www.sciencedirect.com/science/article/pii/S1568494620309455
41. Mahapatra, D., Ge, Z., Sedai, S., Chakravorty, R.: Joint registration and segmentation of xray images using generative adversarial networks. In: Shi, Y., Suk, H., Liu, M. (eds.) Machine Learning in Medical Imaging - 9th International Workshop, MLMI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings. Lecture Notes in Computer Science, vol. 11046, pp. 73–80. Springer (2018)
42. Manzari, O.N., Ahmadabadi, H., Kashiani, H., Shokouhi, S.B., Ayatollahi, A.: Medvit: a robust vision transformer for generalized medical image classification. Computers in Biology and Medicine **157**, 106791 (2023)
43. Matek, C., Schwarz, S., Spiekermann, K., Marr, C.: Human-level recognition of blast cells in acute myeloid leukaemia with convolutional neural networks. Nat. Mach. Intell. **1**(11), 538–544 (2019)

44. Mohamed, M.M.A., Far, B.H., Guaily, A.: An efficient technique for white blood cells nuclei automatic segmentation. In: Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, SMC 2012, Seoul, Korea (South), October 14-17, 2012. pp. 220–225. IEEE (2012)
45. Nikolenko, S.: Synthetic data for deep learning, vol. 174. Springer (2021)
46. Pandey, P., P., P.A., Kyatham, V., Mishra, D., Dastidar, T.R.: Target-independent domain adaptation for wbc classification using generative latent search. IEEE Trans. Medical Imaging **39**(12), 3979–3991 (2020)
47. Rajaraman, S., Antani, S.K., Poostchi, M., Silamut, K., Hossain, M.A., Maude, R.J., Jaeger, S., Thoma, G.R.: Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images. PeerJ **6**, e4568 (Apr 2018). https://doi.org/10.7717/peerj.4568, https://doi.org/10.7717/peerj.4568
48. Rastogi, P., Khanna, K., Singh, V.: Leufeatx: Deep learning-based feature extractor for the diagnosis of acute leukemia from microscopic images of peripheral blood smear. Comput. Biol. Medicine **142**, 105236 (2022), https://doi.org/10.1016/j.compbiomed.2022.105236
49. Rezatofighi, S.H., Soltanian-Zadeh, H.: Automatic recognition of five types of white blood cells in peripheral blood. Comput. Medical Imaging Graph. **35**(4), 333–343 (2011)
50. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, vol. 9351, pp. 234–241. Springer International Publishing, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28, http://link.springer.com/10.1007/978-3-319-24574-4_28, series Title: Lecture Notes in Computer Science
51. Rubin, R., Anzar, S.M., Panthakkan, A., Mansoor, W.: Transforming healthcare: Raabin white blood cell classification with deep vision transformer. In: 2023 6th International Conference on Signal Processing and Information Security (ICSPIS). pp. 212–217 (2023). https://doi.org/10.1109/ICSPIS60075.2023.10344258
52. Saleem, S., Amin, J., Sharif, M., Mallah, G.A., Kadry, S., Gandomi, A.H.: Leukemia segmentation and classification: A comprehensive survey. Comput. Biol. Medicine **150**, 106028 (2022)
53. Salehi, R., Sadafi, A., Gruber, A., Lienemann, P., Navab, N., Albarqouni, S., Marr, C.: Unsupervised cross-domain feature extraction for single blood cell image classification. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention - MICCAI 2022 - 25th International Conference, Singapore, September 18-22, 2022, Proceedings, Part III. Lecture Notes in Computer Science, vol. 13433, pp. 739–748. Springer (2022)
54. Shin, H., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D.J., Summers, R.M.: Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans. Medical Imaging **35**(5), 1285–1298 (2016)
55. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015)
56. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1–9 (2015)

57. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2818–2826 (2016)
58. Tajbakhsh, N., Jeyaseelan, L., Li, Q., Chiang, J.N., Wu, Z., Ding, X.: Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. Medical Image Anal. **63**, 101693 (2020)
59. Tavakoli, S., Ghaffari, A., Kouzehkanan, Z.M., Hosseini, R.: New segmentation and feature extraction algorithm for classification of white blood cells in peripheral smear images. Scientific Reports **11**(1), 19428 (2021)
60. Togaçar, M., Ergen, B., Cömert, Z.: 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), 106810 (2020)
61. Tummala, S., Suresh, A.K.: Few-shot learning using explainable siamese twin network for the automated classification of blood cells. Medical Biol. Eng. Comput. **61**(6), 1549–1563 (2023)
62. Vogado, L.H., Veras, R.M., Araujo, F.H., Silva, R.R., Aires, K.R.: Leukemia diagnosis in blood slides using transfer learning in cnns and svm for classification. Engineering Applications of Artificial Intelligence **72**, 415 – 422 (2018)
63. Waheed, A., Goyal, M., Gupta, D., Khanna, A., Al-Turjman, F., Pinheiro, P.R.: Covidgan: Data augmentation using auxiliary classifier GAN for improved covid-19 detection. IEEE Access **8**, 91916–91923 (2020)
64. Walter, W., Pohlkamp, C., Meggendorfer, M., Nadarajah, N., Kern, W., Haferlach, C., Haferlach, T.: Artificial intelligence in hematological diagnostics: Game changer or gadget? Blood Reviews **58**, 101019 (2023)
65. Wang, J., Lan, C., Liu, C., Ouyang, Y., Qin, T., Lu, W., Chen, Y., Zeng, W., Yu, P.S.: Generalizing to unseen domains: A survey on domain generalization. IEEE Trans. Knowl. Data Eng. **35**(8), 8052–8072 (2023)
66. Wolleb, J., Bieder, F., Sandkühler, R., Cattin, P.C.: Diffusion Models for Medical Anomaly Detection. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. pp. 35–45. Lecture Notes in Computer Science, Springer Nature Switzerland, Cham (2022). https://doi.org/10.1007/978-3-031-16452-1_4
67. Xiang, S., Fu, Y., You, G., Liu, T.: Taking A closer look at synthesis: Fine-grained attribute analysis for person re-identification. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021. pp. 3765–3769. IEEE (2021)
68. Zedda, L.: Wbc-usid (Aug 2024). https://doi.org/10.6084/m9.figshare.26820652, https://figshare.com/articles/software/WBC-USID/26820652/1
69. Zhang, R., Han, X., Lei, Z., Jiang, C., Gul, I., Hu, Q., Zhai, S., Liu, H., Lian, L., Liu, Y., Zhang, Y., Dong, Y., Zhang, C.Y., Lam, T.K., Han, Y., Yu, D., Zhou, J., Qin, P.: Rcmnet: A deep learning model assists CAR-T therapy for leukemia. Comput. Biol. Medicine **150**, 106084 (2022)
70. Zheng, X., Wang, Y., Wang, G., Liu, J.: Fast and robust segmentation of white blood cell images by self-supervised learning. Micron **107**, 55–71 (2018)