HematoFusion: A Weighted Residual-Vision Transformer Ensemble for Automated Classification of Haematologic Disorders in Microscopic Blood Images

Mouna Saadallah, Latefa Oulladji, Farah Ben-Naoum
Evolutionary Engineering and Distributed Information Systems Laboratory, Department of Computer Science, Djillali Liabes University, Sidi Bel Abbes, Algeria
E-mail: mouna.saadallah@univ-sba.dz, latifa.oulladji@univ-sba.dz, farah.bennaoum@univ-sba.dz

Keywords: Medical imaging, neural networks, red blood cell, leukemia, lymphoma

Received: May 27, 2025

Haematologic malignancies pose a significant global challenge, with 1.34 million new cases reported in 2019 and leukemia claiming 311,594 lives in 2020. Early diagnosis of these blood disorders increases survival chances by enabling prompt treatment, yet their complexity and variable cellular morphology hinder accurate detection. Advances in Medical Imaging and AI, particularly Image Classification, offer solutions by analyzing blood samples for subtle morphological patterns. This study advances the field by introducing a novel data set for the classification of red blood cells and using open-source data for the classification of leukemia and lymphoma (each covering 29,363; 16,811; and 1,436 images, respectively). We fine-tuned multiple AI models, including EfficientNetB3, ResNet50V2, and a pretrained Vision Transformer (ViT), and combined their strengths into a weighted ensemble framework. Evaluated across various metrics (including accuracy, precision, recall, etc.), the proposed HematoFusion model excelled, achieving 96% accuracy in the morphology of red blood cells, 99% in Leukemia, and 96% in Lymphoma, surpassing most existing models in terms of accuracy while covering a wider range of haematologic disorders. These findings demonstrate the potential of integrated AI frameworks to improve haematologic diagnostics with precision and reliability.

Povzetek: HematoFusion je uteženi ansambel ResNet50V2 in Vision Transformer, namenjen avtomatski klasifikaciji hematoloških motenj iz mikroskopskih slik. Sistem uporablja nov RBC-nabor podatkov ter odprtokodne nabore levkemije in limfoma ter izboljša zanesljivost diagnostičnega razpoznavanja krvnih celic.

1 Introduction

The collection of blood samples is crucial to understanding diseases, preventing them, and thoroughly providing treatment.

The diagnosis of blood cell diseases hinges significantly on determining the patient's Blood Cell Count (BCC) and observing the appearance of cells under a microscope. It serves as a guide for the pathologist or biologist, providing vital information on diseases that are indicative of quantitative (variations in the number of cells) or qualitative (structural or functional) abnormalities in blood cells [11].

Patients admitted to consultation often suffer haematologic dysfunction (either qualitative or quantitative), some examples of the most common cases requiring medical evaluation are caused either by a decrease in complete blood count, anemia, for instance, sees a decrease in the number of Red Blood Cells (RBCs) or the level of hemoglobin, or an increase/ concentration in the amount of RBCs, as marked in the condition of Erythrocytosis. Other conditions mark a change in the cell's shape and/or size, including: microcyte, macrocyte, echinocytes, codocyte, acanthocyte, spherocyte, and more. White Blood Cells (WBC) and platelet disorders can mainly be described as quantitative, for example: Leukopenia, leukocytosis, neutropenia, lymphocytopenia (WBC), thrombocytosis, or thrombocytopenia (platelets). Most qualitative disorders are of cancer or proliferative disorders, including Leukemia, lymphoma, myeloma (WBC), and Hemophilia (platelets) [9].

The pathologist, along with other medical professionals, depends on studying and examining body tissues to perform diagnostics. The microscope is the main tool used to observe blood cells, which provides a detailed description of them in terms of shape and count. Blood cell observation can be extremely challenging with the naked eye and requires enormous concentration and focus; modern technologies, however, recommend new techniques involving the use of a camera to capture microscopic images that can be exploited for further studies and examination. Some existing solutions like EasyCell ®Assistant, Vision Hema®Assist, include a stand-alone tool using highly costly robots and integrated microscopes that can assist the pathologist to make decisions and save time; nonetheless, due to their high costs and unavailability in public hospi-

tals and laboratories, these solutions cannot be relied upon entirely. This leads us to consider cheaper and more effective innovations emerging in recent years, including Deep Learning (DL) and its various contributions. DL has been widely implemented by researchers in the medical field, and it has given promising results regarding medical imaging (MRI, X-Rays, CT...) [30, 52] and enabled medical professionals to rapidly diagnose and detect abnormalities in the human body without exhausting analysis and observations.

This paper aims to improve classification accuracy for haematologic diseases by leveraging ensemble learning techniques applied to multi-source microscopic datasets, preserving the full spectrum of morphologic variability. The latest DL techniques were exploited, including transfer learning and fine-tuning Convolutional Neural Network (CNN) models and the recently emerging visual transformer (ViT) [28]. The ResNet50V2 and EfficientNetB3 networks were chosen as they were preferable for microscopic image classification, and the latter was suitable for scenarios with limited computing resources. We acquired different sources for our data set that study not only Red Blood Cell disorders but also White Blood Cells (WBC). The CNN and ViT models were separately trained using the completed data set, and the results were later combined to enhance the performance.

A description of our contribution is provided on the following lines:

- A meticulously curated data set for Red Blood Cell Morphology using samples collected in the Anti-Cancer Center in El-Oued, Algeria.
- The base architecture of EfficientNetB3 was used with transfer learning, leveraging pretrained weights from ImageNet. It was additionally fine-tuned for the task of blood cell classification.
- The ResNet50V2 was also integrated and transfer-learned as a base architecture and eventually fine-tuned by adding dense layers and regularization techniques that serve to enhance the model's performance.
- A pretrained ViT model was applied to our data set to classify blood cell images through self-attention mechanisms. The model was fine-tuned by optimizing hyperparameters to improve accuracy.
- A hybrid CNN/ViT model was developed by combining the strengths of CNNs for local feature extraction with those of ViT which captures global features more efficiently.

2 Related work

Pathology and detecting blood disorders require a mass of work and time by a biologist to prepare the blood, test it, and analyze it.

Nevertheless, the emergence of developed technologies, such as deep learning, made things much easier for biologists/ pathologists, as it assists them with the process of analyzing the blood smear and detecting abnormalities in cell type, shape, and aggregation. If done entirely by the pathologist, this step may take hours or even days when necessary, which causes a decline in the health worker's focus and even eyesight.

This urged the need to automate the task to alleviate the pressure on them. Many studies have been conducted to address this problem by exploiting the use of Artificial Intelligence and its diverse techniques.

In its earliest phase, peripheral blood image analysis was inspired by the emerging use of Artificial Intelligence in the medical field and its automation. Kim KS, et al. [2] designed a system that uses a CCD camera attached to the microscope to capture the peripheral images, preprocessing techniques such as edge enhancement and noise removal were applied, and the images were later classified into 15 types of Red Blood Cell abnormalities and 5 normal shapes of White Blood Cells using neural networks. Following that, neural networks, mainly Convolutional Neural Networks, were explored for blood cell image analysis and classification. WBC and its 5 different normal cell shapes (Neutrophils, Lymphocytes, Monocytes, Basophils, Eosinophils) were the easiest to classify and readily available [14] [56] [37]. Classification accuracy reached 96% using a simple neural network that consists of 16 neurons input layers and 10 nodes in a single hidden layer to achieve a minimum error less than 10^{-4} and the output layer with 5 neurons to classify each type [14]. Ali et al. [54] proposed the VGG16-ViT network that uses two online datasets to classify WBC subtypes, achieving excellent precisions of 98.99% and 99.95% on each dataset.

The DenseNet121 model [12] was used by Bozkurt F. [27] on the open-access data set provided by Paul Mooney, available on Kaggle.com [18], reaching an accuracy of 98%. Another Two-Module Deformable CNN and Transfer learning was proposed by Yao Xufeng, et al. [37], whilst the first module initializes the ImageNet [3] characteristic weights, the second module was designated for classification. The authors achieved precisions of 95.7%, 94.5%, and 91.6% for low-resolution and noisy undisclosed data sets, BCCD data set [20], respectively.

Some of the studies, however, focused solely on the classification of one disease. Leukemia is one of the most common blood cancers, leading to growing interest in developing new diagnostic systems for early detection and prevention. In this context, CNNs have gained significant attention due to their efficiency and high accuracy in image-based classification tasks. Areen K. et al. [47] compared in their study multiple CNN-based algorithms (AlexNet, DenseNet, ResNet, and VGG16), employing three datasets (ALL-IDB, ASH ImageBand, and images captured at JUST); reaching an accuracy of 94%. DeepLeukNet proposed by Saeed et al. [53] was conceived to classify Acute Lymphoblastic Leukemia (ALL) subtypes

employing a CNN-based classifier on the ALL-IDB1 and ALL-IDB2 datasets, attaining 99.61% accuracy. Kasim et al. [55] leverages the online datasets ALL-IDB and Munich AML Morphology datasets for multi-class classification of Leukemia subtypes using pretrained CNN architectures and other classification models, including RandomForest, SVM, and Extreme Gradient Boosting. The highest accuracy achieved by this method was of 88%. In recent studies, Vision Transformers (ViTs) have been employed for the classification of Leukemia subtypes. Swain et al. [59] proposed in their research a model based solely on ViTs and classified ALL subtypes. The accuracy of the test set reached 99.67%. A similar approach was implemented by Prasad et al. [51] who attained an overall accuracy of 98.01% for the automatic detection of ALL. Others opted for architectures combining both CNNs and ViTs to further enhance feature extraction. For instance, Tanwar et al. [60] combined in their study the ResNet50 model with the ViT, establishing a dual-stream architecture, reaching an accuracy of 99%.

DL also proved efficient in the classification of other types of cancer such as Lymphoma. Its potential was thoroughly explained by several researchers [58] [35], stressing the application of CNNs and ensemble techniques. Ozgur et al. [49] developed a triple classification system of various Lymphomas: CLL, FL, and MCL, and employed a combination of ML and DL algorithms, reaching precisions of 94%, 92%, and 82%, respectively.

Sickle Cell Anemia and Malaria can be diagnosed by examining the patient's RBCs. Harahap Mawaddah, et al. [29] used a data set that regroups 27.588 images of infected and healthy individuals' RBCs provided by Yasmin M. Kassim et al. [23]. 2 CNN architectures were compared during the classification. LeNet-5 [1] was deemed more precise than DRNet [46] in classifying RBCs affected by Malaria, with accuracies of 95.7% and 95%, respectively. Alzubaidi Laith, et al. [22] introduced a CNN classifying RBC into 3 classes, namely normal, abnormal, and miscellaneous. They used the same network as a feature-extractor, then applied the Error Correcting Output Codes (ECOC) classifier for the classification task, achieving an accuracy of 92.06% In addition to neural networks, Machine Learning techniques were also employed to address the problem of Blood Cell Image Analysis. Aliyu Hajara Abdulkarim, et al. [17] compared Support Vector Machine (SVM) against Deep Learning methods using AlexNet architecture [5]. The dataset used was open-sourced and distinguished 4 types of RBC abnormalities along with their normal shape. The accuracy for the CNN model was relatively weak and couldn't exceed 33%, while the SVM model achieved a perfect 100% on the RBC data set. The latter was deployed with the Radial Basis Function (RBF) default setting; this same network was employed by Syahputra Mohammad Fadly, et al. [15], achieving an accuracy of 83.3% using Canny Edge Detection for preprocessing and feature extraction to classify 3 types of RBC abnormalities.

Label-free identification was also explored by various researchers using an imaging flow cytometer to classify unstained WBCs [19] and optofluidic time-stretch microscopy along with Machine Learning for aggregated platelets detection as well as single platelets and WBC [13].

Visual or Vision Transformers were introduced by Dosovitskiy, et al. [28] in 2020, to exploit transformers in visual applications. Given that image classification is rather a novel concept for transformers, it may take a while to fully develop and exploit ViT in this regard. Compared to ViT, CNN can handle large-scale data sets better and offer excellent results. ViT, however, is known for its understanding of global context and dependencies, although it requires pretraining large amounts of data to achieve comparable results to CNN. [34] Therefore, an ensemble ViT/CNN model can be an excellent approach to incorporate ViT's efficiencies with CNN, this was previously done by Y.Barhoumia, et al. [26] to address another consistent problem of Intra Cranial Hemorrhage Classification. It was also employed by Jiang Zhencun, et al. [32] to diagnose ALL. The ensemble model method used is the weighted-sum model; the output results of the ViT models are multiplied by a coefficient of 0.7, and the output results of the EfficientNet model [21] are multiplied by a coefficient of 0.3. The authors later combined the results to get the final prediction result. The ViT-CNN ensemble model achieved outstanding results with an accuracy of 99.03%, exceeding the models in the literature.

A comparative summary of recent studies on cancer classification using deep learning methods is presented in Supplementary Material: Section 1 (Table S1), which provides the datasets used, classification techniques, number of classes, and accuracy values reported.

3 Methods

3.1 Data acquisition

The data set used for the classification was acquired by combining different sources.

1. The **Chula RBC-12-data set** [33] of RBC blood smear images, which contains a total of 706 smear images describing 13 classes of RBC, and comprising over 20K images of normal and pathological RBC. The images provided were collected at the Oxidation in Red Cell Disorders Research Unit, Chulalongkorn University in 2019, with a DS-Fi2-L3 Nikon microscope used at 1000x magnification. The 13 classes are specified as follows: Normal cell, Macrocyte, Microcyte, Spherocyte, Target cell, Stomatocyte, Ovalocyte, Teardrop, Burr cell, Schistocyte, uncategorized, Hypochromia, Elliptocyte. 2 classes were neglected for the lack of blood smear images.

2. The **ThalassemiaPBS-data set** [40] contains 7108 images of peripheral blood smear images of four thalassemia patients for nine cell types (Elliptocyte, Teardrop, Normal cell, Cigar cell, Stomatocyte, Target

cell, Hypochromia, Spherocyte, Acanthocyte). The images were collected by a clinical pathologist from the Clinical Pathology Laboratory of the Faculty of Medicine, Public Health and Nursing, Universitas Gadjah Mada, Indonesia, using the Olympus CX21 microscope attached with an Optilab advance plus camera with 1000x total magnification.

3. The **RBC-mini data set, Anti-Cancer Center El-Oued, Algeria** [57]: A small data set fragment (mini-batch) provided by the specialized healthcare facility: Anti-Cancer Center in El-Oued, Algeria, that contains a total of 13 blood smear images, regrouping 5 different types of RBC disorders: Burr-cells, ovalocyte, schistocyte, stomatocyte, and teardrop. The blood smear images were captured in May 2024 using an optical microscope with a x1000 magnification. These images were integrated to augment the diversity of the RBC class and mitigate overfitting risks, not to serve as a core data source.

   Table 1 regroups all 3 sources of RBC data sets and lists the size of each data set per type of cell disorder, before and after the application of data augmentation techniques described in Section 3.3.1.
   The total size of the RBC data set is 29,363.

4. The **Raabin-Leukemia data set** [39] is a free-access data set of microscopic images of blood cells, focusing on cases related to Leukemia. 2 experts labeled the cells, and the samples were captured from patients at the Takht-e Tavous laboratory in Tehran, Iran. The Zeiss microscope and LG J3 smartphone camera were used for the imaging.

5. The **Malignant Lymphoma Classification data set** [4] contains a significant number of labeled histopathological images of lymphoma, 3 types of this cancer are covered in this data set: Chronic lymphocytic leukemia (CLL), follicular lymphoma (FL), and mantle cell lymphoma (MCL), through biopsies sectioned and stained with Hematoxylin/Eosin (H+E).

   Tables 2 and 3 present the Lymphoma and Leukemia datasets, respectively, compiled from the Malignant Lymphoma Classification dataset and the Raabin-Leukemia dataset. The tables show the class distribution before and after applying the data augmentation techniques described in Section 3.3.1. The Lymphoma dataset consists of 1,436 images, while the Leukemia dataset contains 16,811 images.

3.2 Image cropping

Figure 1 presents a representative blood smear image from the Chula RBC-12 data set.[33] Each image was manually cropped to focus on individual RBCs and relevant regions of interest. They were subsequently categorized based on specific morphological characteristics. The organization of these classes was performed meticulously to ensure consistency with the referenced files provided by the authors.
The accompanying ”Label” folder within the data set houses a series of files providing detailed annotations for each image, structured in a specific format: the x-coordinate, y-coordinate, and the corresponding RBC type encoded as a numerical value (each class is given a unique value from 1 to 11).
This labeling system facilitates the task of accurate identification and classification of RBCs, thereby serving as a foundation for various haematological studies and the development of automated diagnostic tools.
The same process was replicated on the RBC-mini data set which we collected in collaboration with the Anti-Cancer Center in El-Oued, Algeria. The resulting blood smears were preprocessed and cropped using the OpenCV library as described below. The extracted images were manually labeled under the supervision of specialists at the Center.

1. Load the Image using OpenCV.
2. Preprocess the Image by converting it to grayscale and applying thresholding or edge detection to highlight the cells.
3. Find the cells using contour detection.
4. Extract each cell based on the detected contours and save them as separate image files.

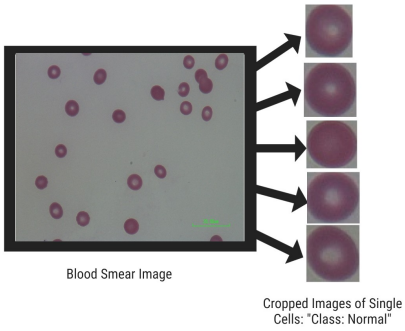


Figure 1: Images of single-cells cropped from one blood smear image

The images provided by the **ThalassemiaPBS-data set** [40] already consisted of single cells; therefore, no further preprocessing was needed. This process was necessary to isolate and classify specific morphological abnormalities.
In contrast, the leukemia [39] and lymphoma [4] data were not cropped as single cells. Instead, the whole blood smear images were retained as input, since the spatial context and global information contained in the whole smear image are all positively contributing towards the classification of leukemia subtypes and malignant lymphomas. These differences in preprocessing reflect the varying nature of the diagnostic tasks and were

Table 1: The complete Red Blood Cells data set description by type of cell and data size, including before and after augmentation

Index	Type of RBC	No. of images [33]	No. of images [40]	No. of images [57]	Total (with augm)
1	Acanthocyte	0	354	0	1432
2	Burr Cell	90	0	10	982
3	Cigar Cell	455	24	0	1893
4	Hypochromia	90	222	0	1284
5	Normal	1812	1426	0	3292
6	Ovalocyte	114	1211	4	3735
7	Schistocyte	108	0	8	453
8	Spherocyte	92	562	0	2640
9	Stomatocyte	49	382	3	1792
10	Target Cell	651	851	0	3912
11	Teardrop	26	2085	6	7948

Table 2: The Lymphoma data set description by type of cell and data size, including before and after augmentation

Category	Subtype	Before	After
Lymphoma	CLL	113	443
	FL	139	526
	MCL	122	467

Table 3: The Leukemia data set description by type of cell and data size, including before and after augmentation

Category	Subtype	Before	After
Leukemia	ALL (L1)	377	1131
	ALL (L2)	3595	3595
	AML (m0)	672	997
	AML (m1)	425	1700
	CLL	1071	3741
	CML	1624	5647

taken into account during the design of the model pipelines.

3.3 Data processing

3.3.1 Data augmentation

Data augmentation is a technique that is essential in image processing. It consists of artificially enhancing the size of a given data set by making changes to the original images. Furthermore, this method presents a solution to improving the model's performance by mitigating common issues like overfitting.

These variations of the existing images generated by the data augmentation techniques provide a more robust data set. These alterations can consist of simple geometric transformations, and color or noise introductions, all designed to make the model's predictions more generalizable and accurate.

In the present study, three primary data augmentation techniques were employed, namely: flipping, which involves mirroring the image horizontally or vertically, rotation, which involves altering the image by turning it by a specified degree, and Gaussian blurring, which can help reduce noise and minor details by adding a Gaussian filter to the image.

When combined, these augmentation techniques allowed us to enrich our data set, all the while relying on additional data preprocessing techniques that will be introduced in the following sections.

3.3.2 Data resizing

Another vital preprocessing technique before training the model is "Resizing". Since our data set is acquired from various sources, it is rather imbalanced, and images come in different sizes and shapes. Therefore, the sizes must be standardized into a uniform squared dimension before feeding the images into the model. This will allow the model to learn efficiently and improve its accuracy. Each model expects a certain target size for the images. ResNet50V2 model, for instance, requires a target size of (224, 224, 3); we were able to apply it using the flow_from_directory() method in Keras. EfficientNetB3, however, expects input images of shape (300, 300, 3) by default, but the model can accept other input shapes as long as the shape is at least 224 × 224 and the number of channels is 3 (RGB); thus, the input size was resized to (224, 224, 3) to reduce computation time and memory usage.

When provided with the target size, Keras uses bilinear interpolation by default for the image resizing operation. The formula specified below is a representation of the process in which the original coordinates are mapped to new ones using interpolation.

$$\text{new}(i', j') = \text{interpolate}\left(\text{orig}\left(i \cdot \frac{H_{\text{orig}}}{H_{\text{tgt}}}, j \cdot \frac{W_{\text{orig}}}{W_{\text{tgt}}}\right)\right) \quad (1)$$

3.3.3 Data rescaling

To ensure uniformity across input data and improve model training, all images were rescaled using appropriate preprocessing techniques depending on the model architecture. To further enhance the CNN-based models' efficiency, we've used "Rescaling", a technique in which the image's range of pixel values is changed to a standard or normalized range.

There are two common rescaling techniques: Standardization and normalization. In our paper, we've opted for the latter, which ensures that various pixel values are used during the model's learning process. The pixels of a given image can be represented as integers in the range of 0 to 255 in the case of an 8-bit image. Rescaling modifies these values into a different range of -1 to 1 or 0 to 1 when using normalization.

Likewise, we've used the flow_from_directory() method to rescale the images by a factor of 1/255 for our training, validation, and test batches. This method uses a form of Min-Max Scaling, where each pixel value is divided by 255. The minimum value (0) in this case maps to 0, and the maximum value (255) in turn maps to 1. Its formula can be defined as follows:

$$X_{\text{scaled}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{2}$$

Meanwhile, for the ViT model, a different preprocessing strategy was implemented to fit its expected input distribution. Mean–standard deviation normalization was applied to standardize the image data and improve the model's convergence. Each image's pixel values were normalized using the following channel-wise means and standard deviations:

- Mean: [0.485, 0.456, 0.406]
- Standard deviation: [0.229, 0.224, 0.225]

This normalization follows the formula:

$$\text{normalized\_pixel}(i, j) = \frac{\text{pixel}(i, j) - \text{mean}}{\text{std\_dev}} \tag{3}$$

Additionally, Supplementary Material: Section 2.2 includes the parameter-level details of the aforementioned data augmentation techniques.

3.4 Proposed solution

In this section, we present the architectures we employed for our blood-cell classification system based on the latest deep-learning techniques. Three state-of-the-art models were explored for this task: EfficientNetB3, ResNet50V2, and Vision Transformer (ViT). To further enhance the classification accuracy, we developed ensemble models combining the strengths of ViTs and CNNs. In training, Transfer Learning was used to fine-tune each of the cited architectures, and the hyperparameters were optimized to suit the specifics of the corresponding data set.

The choice of models and more specific details are explained later in the section.

3.4.1 EfficientNetB3

A member of the EfficientNet family that was first introduced in May 2019 by [21]. This architecture was chosen due to its superior performance in feature extraction and its ability to balance computational efficiency with high accuracy, making it well-suited for tasks like blood cell classification.

EfficientNets are developed based on AutoML and compound scaling. The authors first used the AutoML MNAS Mobile framework to develop a baseline network, which they named EfficientNetB0, the first of the EfficientNet family. They then used the compound scaling method to scale up and obtain the series from B1 to B7.

The architectures achieved higher accuracy and efficiency despite being smaller and, thus faster than other models.

In our paper, we have opted for the B3 version which gave promising initial results, additional layers were added to adapt the model for blood cell classification.

We additionally adjusted key hyperparameters meticulously during training, such as learning rate, batch size, and dropout rate.

Figure 2 shows the architecture of the EfficientNetB3 base model that we have adopted for our specific classification task. The architecture was created using diagrams.net (formerly known as draw.io). [31]

The model is first fed microscopic images resized to (300, 300, 3) and processed through its pretrained backbone. The default fully connected classification head of EfficientNetB3 had been removed since it is only specific to the ImageNet data set it was trained on (containing 1000 classes), which allowed us to add the custom classification head tailored to our data set.

Three versions of the same architecture were used, each with a modified softmax layer to adapt to our three different data sets: *(1)* for RBC classification, 11 classes, *(2)* for Leukemia classification, 6 classes, *(3)* for Lymphoma classification, 3 classes.

The EfficientNetB3 backbone acts as a feature extractor, extracting spatial and hierarchical features that are later fed to the added layers for learning. The first five layers are frozen to prevent the weights from being updated during training. This helps to adapt the deeper layers to our data set; whereas freezing more layers could have resulted in under-fitting since the data set has unique characteristics that are significantly different from the original ImageNet data set. Deeper layers of the EfficientNetB3 backbone are unfrozen to enable the model to capture more patterns. (eg: cell morphology, staining patterns, etc).

This version of the model expects a (300, 300, 3) input shape by default, we have resized the input size to (224, 224, 3) to speed up training and reduce memory usage due to limited resources and the size of our data set which is

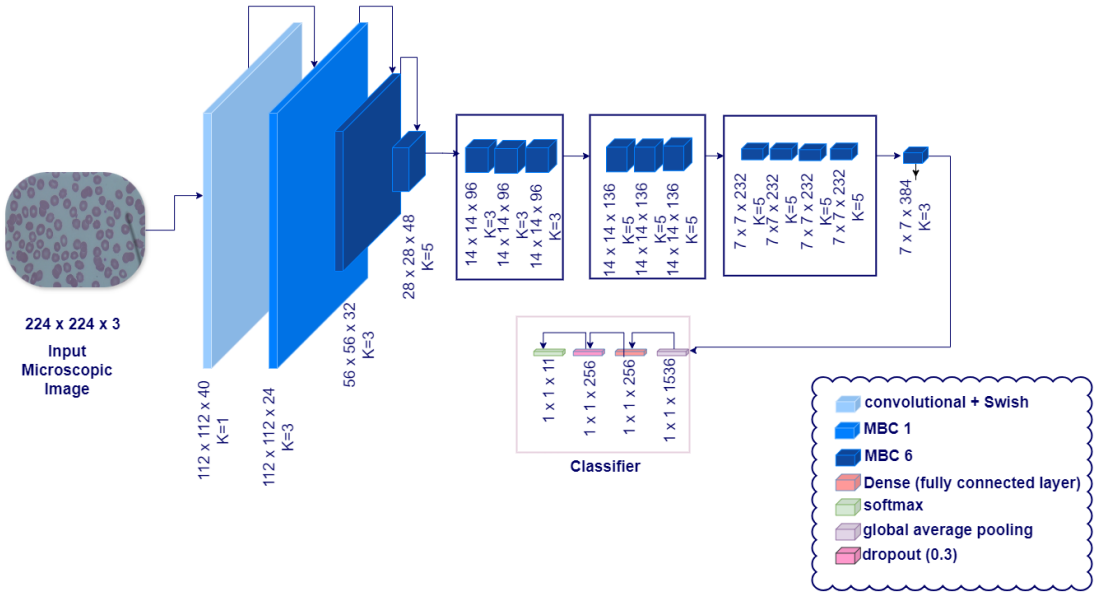


Figure 2: EfficientNetB3 model for blood cell classification: The model processes 300x300x3 microscopic images through convolutional layers with Swish activation, followed by mobile inverted bottleneck blocks (MBConv 1 and 6). The first 5 layers are frozen, with fine-tuned deeper layers. A custom classification head is added for task-specific classification

rather small.

A dropout layer is added after the global average pooling (GAP) to reduce the overfitting that we have been subjected to due to the depth of the network against the small size of the data set. A dropout rate of 0.3 was employed, thus deactivating 30% of neurons during training. This prevents the model from relying on these specific neurons.

A fully connected layer with 256 units using the ReLU activation function is added, serving to learn complex representations.

The classification head is completed with the final output dense layer. The number of units corresponds to the number of classes in each of our 3 data sets. The softmax activation function is used to specify the multi-class classification.

3.4.2 ResNet50V2

Deep convolutional neural networks have contributed to the Image-Classification field significantly, providing a robust platform to researchers ever since the emergence of the first-ever deep neural network; LeNet in 1998. Later on, in 2012, the idea of Dropout was presented, allowing the model to avoid overfitting.

Researchers next focused solely on adding more convolutional layers to increase the depth of the model, and thus, its efficiency. However, the idea of simply stacking up layers didn't benefit researchers as it introduced a whole new issue, the accuracy degradation, which unexpectedly wasn't due to overfitting, rather, it was caused by the Vanishing Gradient Effect. [6]

Residual Neural Networks to the rescue! In 2015, ResNet152, the first of the ResNet family was introduced. It consists essentially of modularized architectures that stack building blocks of the same connecting shape, called short-cut connections, that skip one or more layers. [10] These connections in ResNet work by performing identity mapping, the outputs of this mapping are added to those of the stacked layers as illustrated in Figure 3.

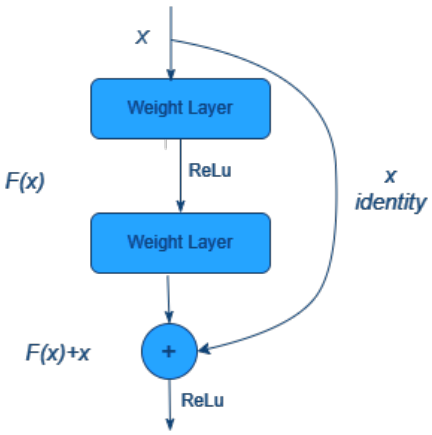


Figure 3: Residual learning - a building block

ResNet50V2 is a residual neural network variant that employs skip connections to prevent vanishing gradients during back-propagation; This ensures efficiency in learning complex features present in microscopic blood cell images.

Figure 4 presents the architecture of the ResNet50V2

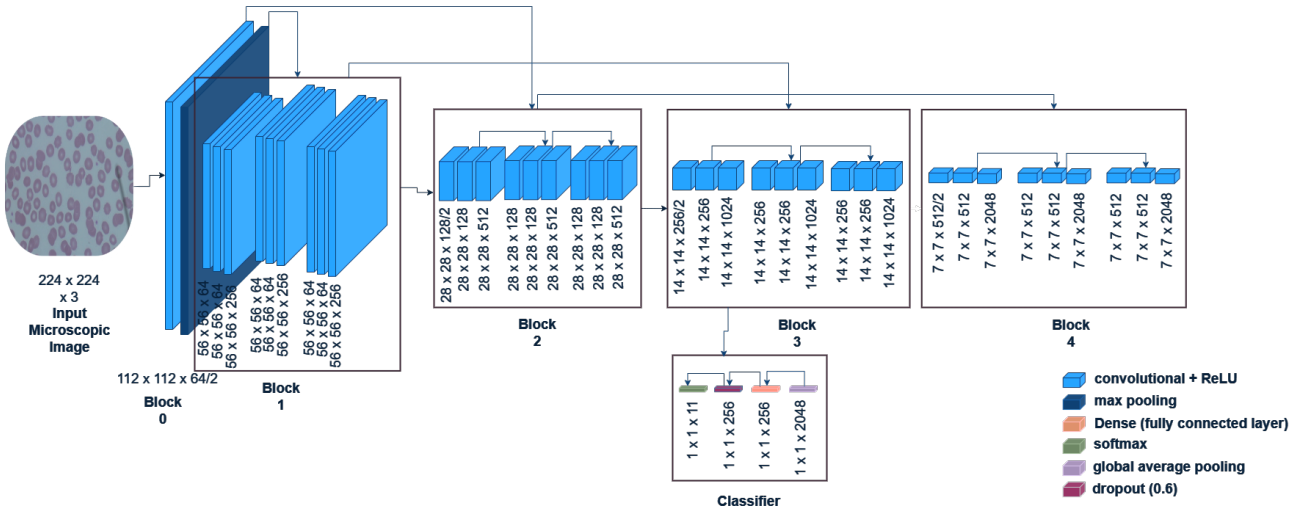


Figure 4: ResNet-50v2 model for blood cell classification: The model processes 224 × 224 × 3 microscopic images through a series of convolutional layers with ReLU activation. It consists of four main blocks with residual connections and employs bottleneck blocks (1x1, 3x3, 1x1 convolutions). A global average pooling layer is added, followed by a fully connected classification head, and a softmax activation for predicting blood cell classes. Key components include skip connections, dropout (0.6), and task-specific fine-tuning

base model that we have employed for our classification. Similarly, the model was also designed using the diagrams.net tool.
The model is fed a microscopic image of size (224,224,3) that is previously preprocessed and normalized.
It consists of 50 layers, focusing more on improved gradient flow and training stability by introducing pre-activation residual blocks and applying batch-normalization and activation (ReLU) before convolutions.
The network's initial block captures low-level features such as edges, textures, and patterns by implementing convolutional and pooling layers, followed by 4 residual blocks, with skip connections to prevent vanishing gradient problems. Higher-level features are extracted using down-sampling (strides).
The final output of these blocks is passed through a global average pooling layer (GAP) to reduce the feature map to a 1D vector; A fully connected layer and a softmax classifier are added.

The base model acts as a feature extractor, and the custom layers act as a task-specific classifier, tailored to blood-cell classification. Similar to the EfficientNet, the model was transfer-learned by freezing the first 5 layers. This prevents overfitting as the learning focuses on the deep layers, and a dropout layer is also added to ensure the model does not memorize the training data. Preceded by a 256 units dense layer that allows the model to combine the learned features to improve classification and the classifier ends with a dense layer that has the same number of neurons as the classification classes: (1) for RBC classification, 11 classes, (2) for Leukemia classification, 6 classes, (3) for Lymphoma classification, 3 classes.

3.4.3 Experimental hyperparameters

Table 4 presents a breakdown of the hyperparameters and setup used in the experiments based on the training pipeline using Keras/Tensorflow, along with their purposes, as well as the strategies employed to transfer-learn and fine-tune the models and achieve the best accuracies possible.

The EfficientNetB3 and ResNet50V2 models were both trained using the same hyperparameters detailed in Table 4.

3.4.4 Experimental environment

Hardware: The experiments for all 3 models were conducted by Google Colab; It typically consists of NVIDIA Tesla GPUs.

Software: *Platform:* Google Colab, a hosted Jupyter Notebook environment.
Framework(s): Tensorflow v2.18.0 and Keras v3.6.0 were used to develop, train, and evaluate the 3 models.
Python: Version (3.10).
Libraries: matplotLib, numpy, PIL, joblib, and more were used to preprocess, analyze, and visualize data.

Storage: The 3 data sets were preprocessed and split into Training, Validation, and Test data sets, each stored in Google Drive, which is mounted to the Colab environment for access. The detailed data split strategy, including training, validation, and testing partitions, is provided in Supplementary Material: Section 2.1.

Table 4: Experimental hyperparameters for training the CNN models and their purposes

Hyperparameter	Value	Purpose
Optimizer	Adam (LR=10^{-4})	A low learning rate is employed to fine-tune pre-trained layers.
Loss function	Categorical crossentropy	Used for multi-class classification.
Metrics	Accuracy	To monitor the number of correctly classified instances during training.
Steps-per-epoch	ResNet50V2: 20, EfficientNetB3: 200	The number of training batches processed per epoch.
Validation steps	ResNet50V2: 10, EfficientNetB3: 316	The number of validation batches processed per validation step.
Epochs	ResNet50V2: 300, EfficientNetB3: 10	Specifies the training schedule which allows gradual convergence.

3.5 ViT model

Visual or Vision Transformers (ViT) is a novel approach introduced by Dosovitskiy et al. [28]. It uses the concept of transformers designed specifically for visual applications and image classification tasks in particular.
When using the transformer blocks in ViT, the multi-head attention mechanism is applied to integrate global context efficiently and learn high-level features. [42]
Following the success of NLP transformers [16], Dosovitskiy et al. were inspired to develop a new attention-based class of models that can be exploited in Computer Vision.
Compared to NLP transformers, ViT only uses the encoder attention branch, neglecting the decoder attention branch, whilst word tokens are replaced by image patches.
In a normal CNN, the entire image is taken as input, whereas in ViT, the image is first divided into equal-sized patches, which are passed through some linear layers; the outputs of this layer are known as patch embeddings. Between these embeddings, we have the position embeddings, which serve to provide the model with some positional information regarding the sequence of these patches. Afterward, another learnable token is added to the position embedding for image classification purposes.

Figure 5 presents the architecture of the ViT model we have employed for our blood cell classification task. Prior to the training phase, the data was first prepared and processed to fit the model's requirements and expected input. The data set was initially split into training, validation, and test sets and stored in specific folders. The ImageFolder utility was used to load the images and associate them with their corresponding classes based on the folder names provided. The images were later resized to fit the shape expected by the ViT model: 224×224. Further normalization was implemented to standardize the image data, and make it more suitable for the model (See Section 3.3.3).

Similarly to the CNN architecture, three versions were implemented for each data set: *(1)* for RBC classification, 11 classes, *(2)* for Leukemia classification, 6 classes, *(3)* for Lymphoma classification, 3 classes.

The pretrained backbone uses the google/vit-base-patch16-224-in21K model from the Hugging Face library [25] as a feature extractor. The model was trained on the ImageNet-21K data set [8]. It was fine-tuned to adapt to the blood cell classification task, where the number of labels was defined as the number of classes in the data set as mentioned previously in this section.

The transformer encoder depicted in Figure 5 is first provided with the embedded patches (Patch-embedding/ Position-embedding). The input image is divided into fixed-size patches of 16x16. We next apply a linear projection on the flattened patches to form a fixed-dimensional vector. Unlike CNNs, Transformers require position embeddings to learn and capture the input's order of sequence [38]. It serves to improve accuracy and encode the spatial information of the patches.

The combined embedded patches are fed into the Transformer Encoder to go through a series of L layers, each including a list of components, as follows:

1. *Multi-head self-attention* is a mechanism that enables the model to learn global patterns by splitting the process of self-attention into multiple heads, where each head focuses on the interaction between patch-embeddings differently.[16] The attention calculations are eventually merged to give a more global score.

2. The *output* of the Multi-head attention is added to the input of the next component by a *skip connection* (residual connection) after normalization. As explained earlier, residual connections are added to prevent the vanishing gradient during training.

3. To further enhance the model's learning through patch-embeddings, a *feed-forward network (FFN)*

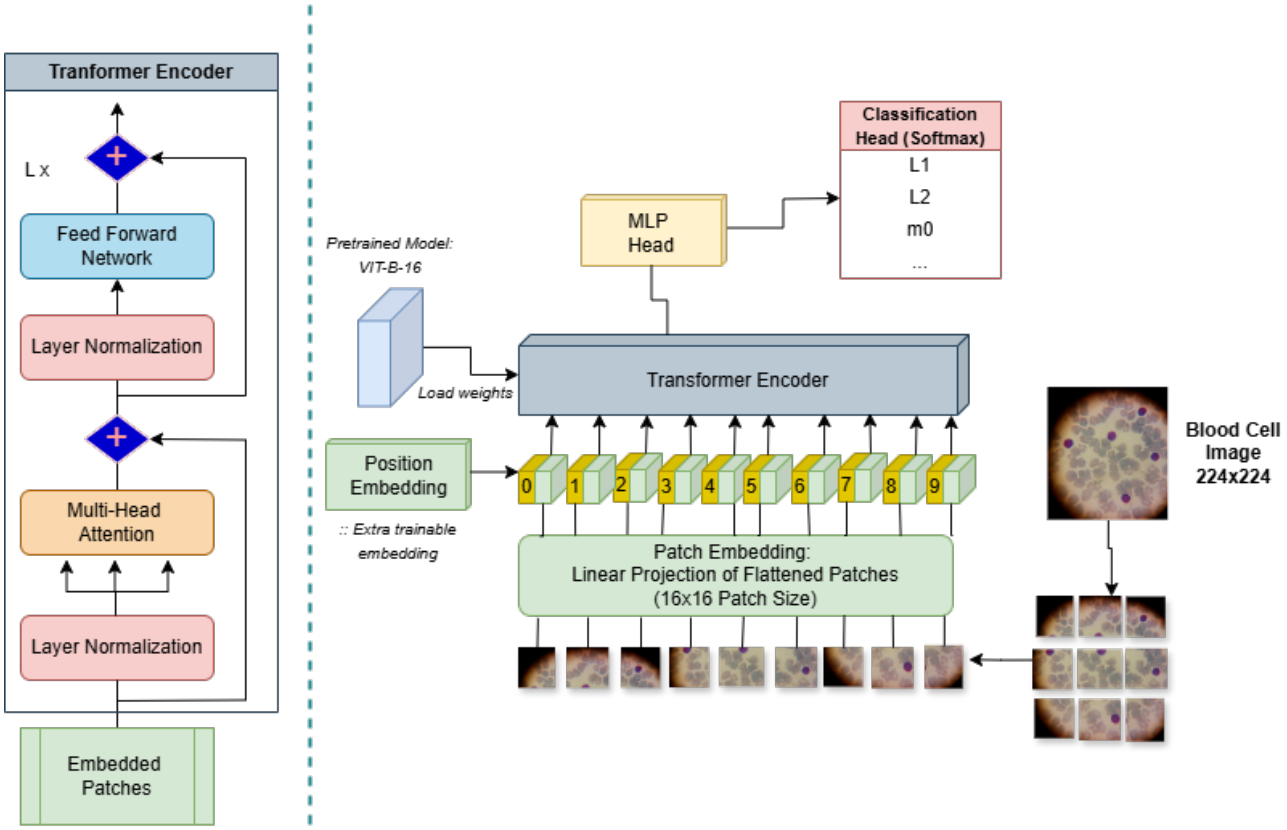


Figure 5: Vision Transformer (ViT) Architecture for blood cell classification: The model processes 224×224 microscopic images through patch embeddings and position encoding, which are later fed to the transformer encoder with loaded weights from the pretrained ViT-B-16 in 21K model. After passing through the transformer encoder, the embeddings are used as the input to the classification head (MLP + Softmax)

[48] is fed to the normalized output of the Multi-head attention, which consists of fully connected layers with a GeLU activation in between. This allows the model to capture local transformations.

4. Similarly to the Multi-head self-attention block, the *FFN* is normalized and added to the *residual connection*.

The output of the Transformer Encoder is a sequence of embedding, enriched with local and global contextual information, independently for each patch.
After passing through the Transformer Encoder, the embedding corresponding to the special classification token (cls) is used as the input to the classification head that consists of a Multi-Perceptron head (MLP), and a softmax classification head.
The MLP takes the output of the Transformer Encoder and feeds it into a series of fully connected layers to prepare the data for the softmax classification head that processes it to the desired classes.

3.5.1 Experimental hyperparameters

Table 5 outlines the hyperparameters used for training a ViT model using the backbone google/vit-base-patch16-224-in21k, along with their respective values; detailing the batch size, the learning rate, and the optimizer employed. The OneCycleLr Scheduler was used as a strategy to vary the learning rate during training; each cycle uses a maximum of 10^{-3} as a learning rate. Other parameters include the CrossEntropyLoss function and a total of 10 epochs (624 batches per epoch) to train the model.

Table 5: Experimental hyperparameters for training the ViT model

Hyperparameter	**Value**
Batch size	32
Learning rate	10^{-4} (initial)
Optimizer	AdamW
Scheduler	OneCycleLR
Scheduler Max lr	10^{-3}
Number of epochs	10
Loss Function	CrossEntropyLoss
Model backbone	google/vit-base-patch16-224-in21k

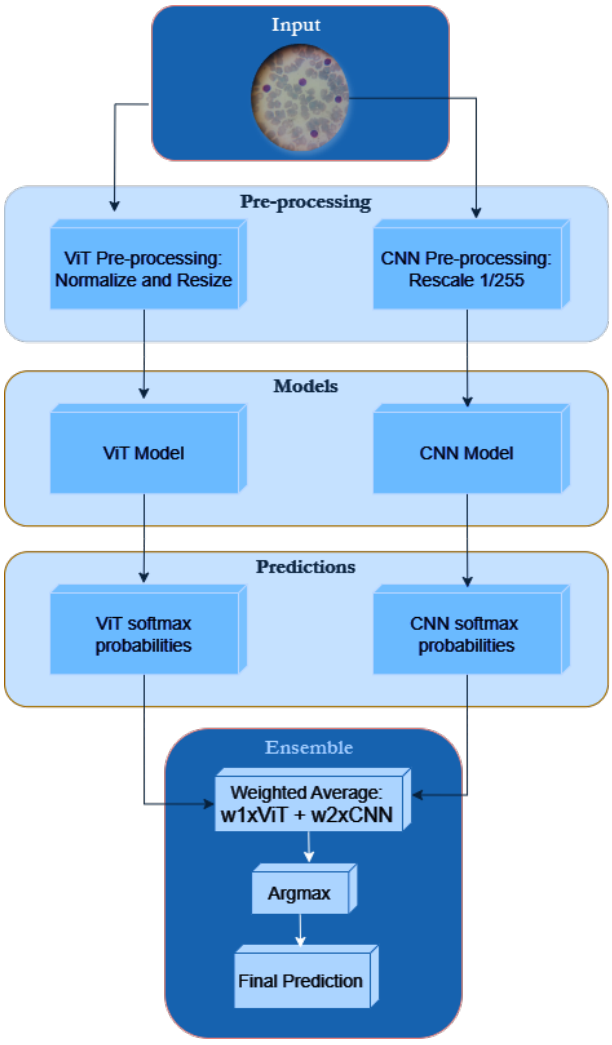


Figure 6: Workflow of the weighted average ensemble method: The input is preprocessed to fit the ViT and the CNN models, and the predictions of the models are later combined using the weighted average approach to generate the final prediction

3.5.2 ViT-CNN ensemble model

To further enhance the performance of our models, an ensemble method was introduced to seek the opinions of several models and combine them to achieve highly accurate classifications than those of the raw models when trained separately [44] [50]. Through our experiments, we have observed the superiority of residual networks in training and efficient learning, while the ViT model performed better in certain instances, focusing more on learning complex features. Thus, we incorporated in our methodology a dual-architecture ensemble, combining the residual network's efficiency with the high precision obtained by the ViT.

Figure 6 presents the flowchart of the ResNet-ViT ensemble model that we have implemented.
The weighted-average ensemble method was selected after experimenting with the most prevalent methods in image classification tasks, namely, maximum voting, the averaging method, and the weighted sum. In the weighted-average method, the models are assigned different weights after training, defining the importance of each model for prediction.

The weighted-average ensemble combines predictions from a CNN (M_1) and a Vision Transformer (M_2), with output probabilities for class c denoted as $P_1(c)$ and $P_2(c)$, respectively, obtained via the softmax function to ensure $\sum_c P_i(c) = 1$ for $i = 1, 2$. The weights w_1 and w_2 are assigned to M_1 and M_2 based on validation performance. The ensemble probability for class c is computed as:

$$P_{\text{ensemble}}(c) = \frac{w_1 P_1(c) + w_2 P_2(c)}{w_1 + w_2} \tag{4}$$

The final class prediction is determined by selecting the class with the highest ensemble probability:

$$\hat{c} = \arg\max_c P_{\text{ensemble}}(c) \tag{5}$$

Preprocessing: The input fed to the already trained models is first preprocessed; each model is preprocessed differently. The ViT uses normalization with mean and std, while the CNN uses simple rescaling (1/255). The models are later loaded to make predictions, and both models output probabilities for the different classes; we used the softmax function to ensure that they sum up to 1.

Weight Selection: Weights w_1 and w_2 were determined through a grid search over predefined pairs, specifically $[(0.3, 0.7), (0.4, 0.6)]$, where each pair sums to 1 to maintain normalized probabilities. The grid search evaluated each weight combination on a validation subset using classification accuracy as the performance metric. The pair that achieved the highest accuracy was selected. Further details about the ensemble weight selection and performance across datasets are provided in the Supplementary Material: Section 3.

4 Results

This section provides an in-depth analysis of the results obtained from our experiments. First, we explore the performance of our individual models, using the insights present in the confusion matrix and focusing on metrics such as: (1) accuracy, (2) precision, (3) recall, (4) F1-score, (5) Cohen kappa, and (6) AUC scores, followed by an evaluation of the ensemble model, HematoFusion. The evaluation is conducted across the three data sets we have introduced in earlier sections, and the results are eventually interpreted in the context of existing literature.

The *Accuracy* is calculated by measuring the number of predicted cases. When achieved, a high accuracy means the overall performance of the model is good. However, in the case of imbalanced data sets, high accuracy can be misleading, and other metrics are necessary to further evaluate the model.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{6}$$

The *Precision* is calculated by measuring the number of correctly predicted positive cases. A high precision is achieved only if most of the positive cases are correctly predicted. [45]

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{7}$$

The *Recall*, also known as *Sensitivity* or *True Positive Rate* measures whether all relevant cases of the data set were correctly predicted.[45]

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

To address the accuracy's shortcomings in handling imbalanced data sets, which is the case in our paper, the *F1-score* was introduced for balanced evaluations, combining precision and recall in one metric. The F1-score eventually only achieves a high accuracy when both precision and recall are high. [43]

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}$$

The *Cohen Kappa* was introduced as a statistical measure of the agreement between the predicted labels and their actual values. If κ=1, the perfect agreement is achieved, if κ=0, the agreement is random, and if κ<0, it means the model achieved more than a random agreement. [36]

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{10}$$

where:

P_o = Observed agreement (accuracy)

P_e = Expected agreement based on chance

$$P_e = \sum_{i=1}^{k} \frac{(A_i \cdot B_i)}{N^2}$$

The *Area Under the Curve (AUC)* score, specifically the Area Under the Receiver Operating Characteristic (ROC) curve [24], evaluates a model's ability to discriminate between classes at various thresholds of classification. An AUC score approaching 1 indicates high discriminative capability, particularly useful for unbalanced datasets that are common in blood cell classification, where accuracy becomes misleading based on class difference.

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d\text{FPR} \tag{11}$$

where:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \text{ (True Positive Rate or Sensitivity)}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \text{ (False Positive Rate).}$$

4.1 Classification results

To analyze the model's performance during training, the accuracy and loss were both monitored and visualized through the training curves over successive epochs.
Figures S5, S6, and S7 (Supplementary Material Section 7) depict the training curves for the RBC Morphology, Leukemia, and Lymphoma data sets, respectively.

Additionally, for further visual evaluation of the classification performance, confusion matrices were computed on the test set, showing the number of accurate and inaccurate predictions of instances, namely: True Negative (TN), True Positive (TP), False Negative (FN), and False Positive (FP).
The confusion matrices generated for the RBC Morphology, Leukemia, and Lymphoma datasets are displayed, respectively, in Figures 7, 8, and 9.

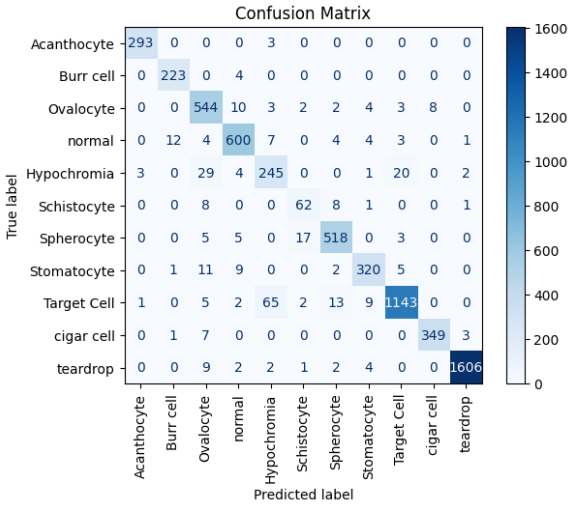
These confusion matrices were used to compute the quantitative metrics for a more specific evaluation.
Detailed tables presenting per-class performance metrics (precision, recall, and F1-score, kappa) for each model and dataset combination are provided in the Supplementary Material: Section 4, Tables S3–S5.

The results for the classification of RBC Morphology, Leukemia, Lymphoma, the Ensemble model, HematoFusion are summarized in Tables 6, 7, and 8, respectively. To test model stability and robustness over sets, we conducted a bootstrapping analysis. This technique provides us with an estimate of results' variability and increases the validity of our performance claims over single-run statistics. The detailed bootstrapping results with distribution plots and summary statistics are presented in Supplementary Material: Section 6 (Table S7 and Figures S2–S4).
Furthermore, we have calculated and included the AUC scores (Table S6) and ROC curves (Figure S1) for our individual models across each dataset, as shown in Supplementary Material: Section 5.

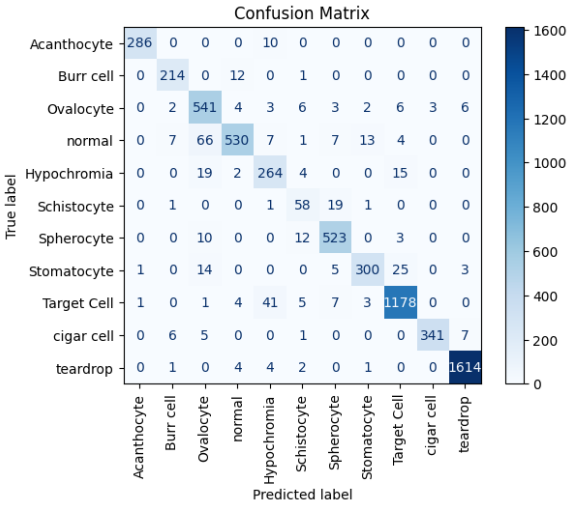
5 Discussion

5.1 Interpretation of results

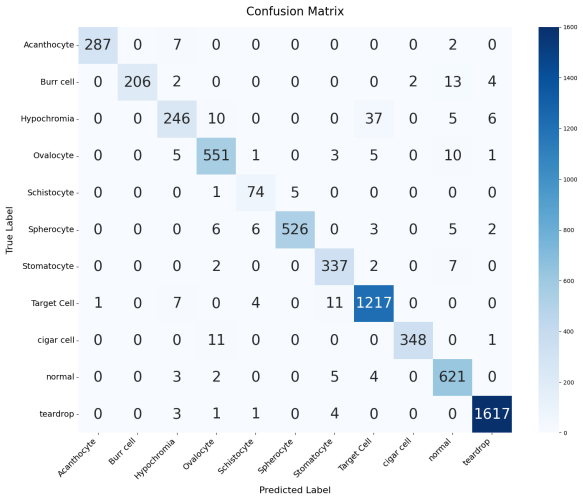
The convergence of the ResNet50V2 model illustrates a steady reduction in training loss, and the accuracy becomes stable after reaching a certain number of epochs. The ViT model has demonstrated higher fluctuation in both accuracy and loss during training.



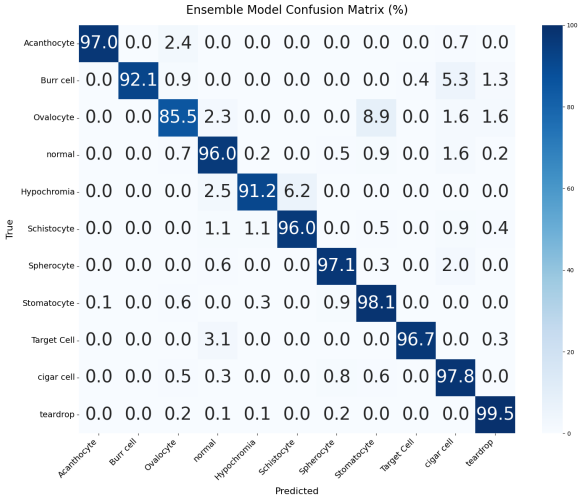
(a) EfficientNetB3 model



(b) ResNet50V2 model



(c) Pretrained ViT model



(d) HematoFusion model

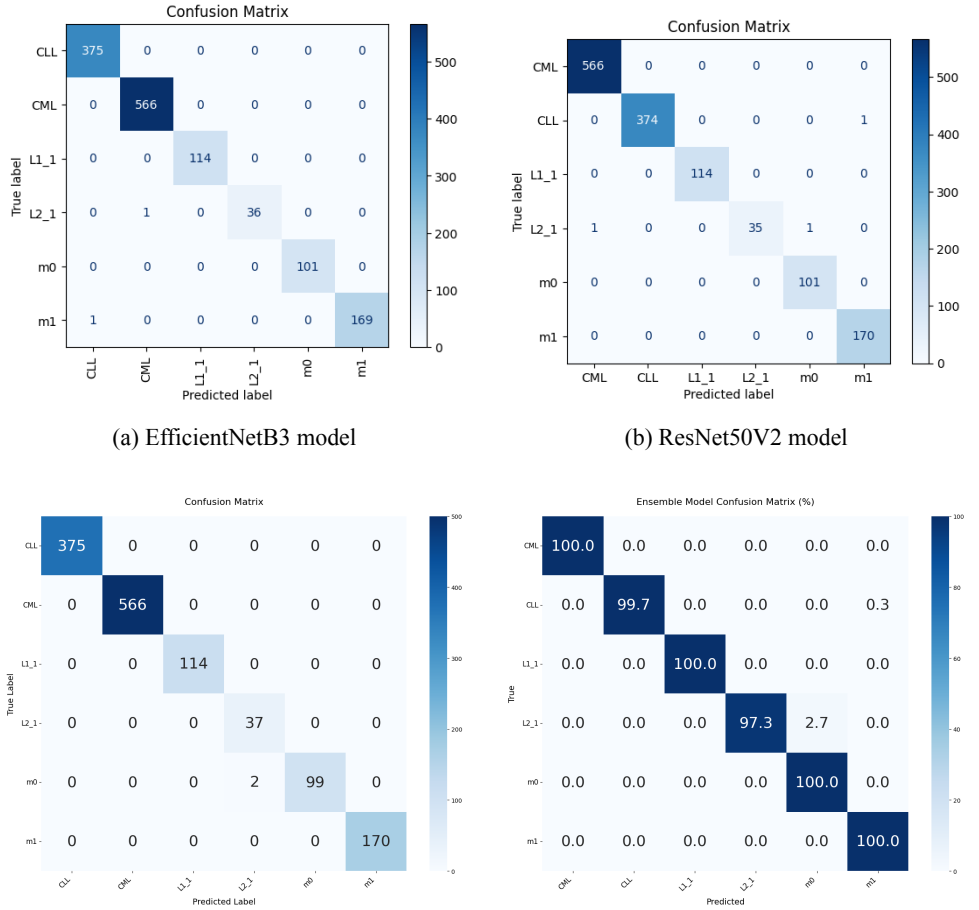
Figure 7: Confusion matrices for the classification performance of the four models on the RBC data set: (a) EfficientNetB3 model, (b) ResNet50V2 model, (c) pretrained ViT model, (d) HematoFusion model combining the ViT and ResNet50V2 models.

Table 6: RBC Morphology classification results across the three individual models with detailed metrics for evaluation

Model	**Model Performance**						
	Train Acc	**Val Acc**	**Test Acc**	**Kappa**	**Recall**	**F1-score**	**Precision**
EfficientNetB3	0.99	0.91	0.97	0.93	0.92	0.92	0.92
ResNet50V2	0.98	0.98	0.92	0.92	0.93	0.93	0.93
ViT	0.98	0.94	0.96	0.96	0.94	0.94	0.94

Table 7: Leukemia classification results across the three individual models with detailed metrics for evaluation

Model	**Model Performance**						
	Train Acc	**Val Acc**	**Test Acc**	**Kappa**	**Recall**	**F1-score**	**Precision**
EfficientNetB3	1.0	1.0	1.0	0.99	0.99	0.99	0.99
ResNet50V2	1.0	1.0	1.0	0.99	1.0	1.0	1.0
ViT	0.99	0.99	0.99	0.99	1.0	1.0	1.0



(a) EfficientNetB3 model (b) ResNet50V2 model

(c) Pretrained ViT model (d) HematoFusion model

Figure 8: Confusion matrices for the classification performance of the four models on the Leukemia data set: (a) EfficientNetB3 model, (b) ResNet50V2 model, (c) pretrained ViT model, (d) HematoFusion model combining the ViT and ResNet50V2 models.

Table 8: Lymphoma classification results across the three individual models with detailed metrics for evaluation

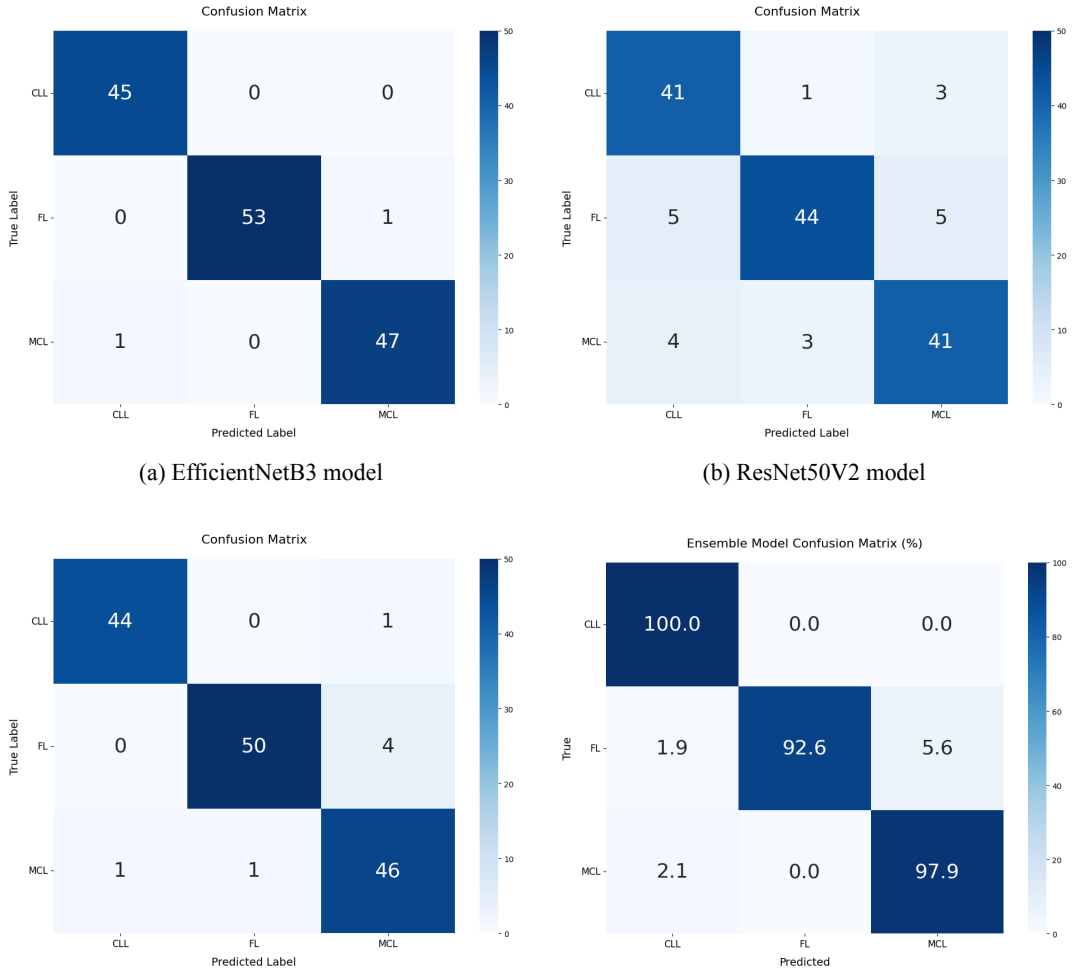
Model	Model Performance						
	Train Acc	Val Acc	Test Acc	Kappa	Recall	F1-score	Precision
EfficientNetB3	1.0	0.99	0.99	0.97	0.99	0.99	0.99
ResNet50V2	1.0	0.91	0.96	0.91	0.96	0.96	0.96
ViT	0.98	0.98	0.95	0.92	0.98	0.98	0.98

Table 9: HematoFusion ensemble model classification results across the three datasets, showing detailed evaluation metrics

Dataset	Ensemble Model Performance				
	Best Acc	Kappa	Recall	F1-score	Precision
RBC	0.96	0.94	0.97	0.97	0.97
Leukemia	0.99	0.99	1.00	1.00	1.00
Lymphoma	0.96	0.95	0.97	0.97	0.97

When comparing the True Positives of the proposed HematoFusion model with the individual models, we can clearly observe an increase in the rates of correctly classified cases and a decrease in the misclassification rates.

The individual models struggled with predicting the Hypochromia class. The ensemble model, on the other hand, exhibited a stronger ability to recognize this class. Whereas, Acanthocyte and Teardrop were easier to identify, owing to their distinguishable shapes, which was reflected in the high number of TP.



(a) EfficientNetB3 model (b) ResNet50V2 model

(c) Pretrained ViT model (d) HematoFusion model

Figure 9: Confusion matrices for the classification performance of the four models on the Lymphoma data set: (a) EfficientNetB3 model, (b) ResNet50V2 model, (c) pretrained ViT model, (d) HematoFusion model combining the ViT and ResNet50V2 models.

In Table 6, a slight overfitting is observed due to the class imbalance; thus, the accuracy can be misleading for accurate measurement of the performance. This, however, was addressed with the use of precision, recall, and F1-score. Although EfficientNetB3 was slightly better on test accuracy (0.97), the ViT model outperformed in stronger and more descriptive metrics, like Kappa score, precision, recall, and F1-score, which indicate more balanced performance with class imbalance. Therefore, ViT is graded as the overall best performer on the RBC morphology dataset. Table 7 shows more consistent results on the Leukemia data set across all models, achieving perfect classification, which indicates better generalizations.

Both ResNet50V2 and EfficientNetB3 achieved comparable top-tier performances on the Leukemia dataset, with identical test accuracy, precision, recall, and F1-score, and minor variations in other evaluation metrics. EfficientNetB3, alternatively, outperforms the other models on the Lymphoma classification (Table 8), reaching almost perfect accuracies.

The Ensemble model, HematoFusion, demonstrates more uniform results across all data sets in terms of all evaluating metrics, mitigating the issues with the class imbalance, as evidenced by its performances, leveraging the strengths of both the ViT and ResNet50V2 models that struggled with some classes.

The precision improved by 4% in the RBC data set and reached a perfect 100% for Leukemia classification, averaging the performances of the individual models on the Lymphoma data set with a precision of 97% on the test set. Despite the strong performance of our proposed solution, further improvement could be implemented to help the model generalize better and address the issue of class imbalance efficiently.

5.2 Comparative study

Table 9 presents a breakdown of the performance of the proposed solution across all three datasets, outlining the accuracy and precision of the model when compared to the lit-

Table 10: Comparative results of the proposed solution and the literature across different metrics for each data set

Dataset	HematoFusion		Literature	
	Accuracy	**Precision**	**Accuracy**	**Precision**
RBC	**0.96**	**0.97**	0.98	0.97
Leukemia	**0.99**	**1.0**	0.99	0.99
Lymphoma	**0.96**	**0.97**	0.96	0.96

erature.
In a bid to substantiate the efficiency of our proposed solution, we evaluated it against the following models:

1. *Literature [7] RBC Classification:* The authors presented a Maximum Voting based Ensemble Model to classify Dacrocyte (Teardrop), Schistocyte, and Elliptocyte cells (Cigar cells) in Iron Deficiency Anemia. The average classification precision and accuracy of the latter reached a maximum of 97% and 98%, respectively; While both models achieved the same precision of 97%, the model in the Literature reported a slightly higher accuracy (98%) compared to HematoFusion's 96%. Nonetheless, it's worth noting that our data set comprises 11 classes against the 3 classes studied in this article.

2. *Literature [32] Leukemia Classification:* The authors proposed a ViT-CNN Ensemble Model for the diagnosis of Acute Lymphoblastic Leukemia (ALL), which is one of the 6 classes that we analyzed in our paper. Compared to the model in the Literature, which achieved 99% accuracy and 99% precision on the Leukemia dataset, HematoFusion matched the accuracy (99%) but outperformed in precision, achieving a perfect 100%.

3. *Literature [41] Lymphoma Classification:* Malignant Lymphoma (ML) was addressed in this paper, which is among the 3 classes that appear in our Lymphoma data set.
   The proposed hybrid model used the combined features of 3 deep learning networks, namely, MobileNet-VGG16, VGG16-AlexNet, and MobileNet-AlexNet, to classify the models by the XGBoost and DT algorithms, reaching an average accuracy and precision of 96%.

An extended version of this comparison, covering a broader range of SOTA models and datasets, is provided in Supplementary Material: Section 8 (Table S8).
Overall, our proposed HematoFusion Ensemble model achieved a reliable performance across the 3 data sets, despite the imbalanced data set and the high number of classes in the case of RBC Morphology Classification.

5.3 Limitations

Although the reported results show high precision, reaching up to 99%, this should be interpreted with caution due to known issues like *dataset imbalance*. As identified previously, some classes were underrepresented, and this could result in biased learning as well as overfitting. To mitigate this, data augmentation techniques were employed (as outlined in Supplementary Section 2.2), and performance was monitored across a variety of metrics (precision, recall, F1-score, Cohen's Kappa, and AUC scores) rather than simply accuracy. However, we are aware that the lack of external validation data limits generalizability. Although high-performance metrics are presented, the models have not been prospectively validated within a real clinical workflow. Their incorporation into clinical decision-making would require extensive regulatory testing and interpretability evaluation. Additionally, while conventional regularization techniques such as dropout and data augmentation were applied to address overfitting, we recognize the need for more advanced strategies. Future work will explore class imbalance mitigation techniques such as SMOTE, GAN-based synthetic image generation, and uncertainty-aware training beyond testing on independent cohorts, to further assess the robustness of the model in actual clinical settings. Furthermore, we intend to conduct ablation studies on ensemble weight parameters and data augmentation strategies to evaluate their individual contributions.

6 Conclusion

In this study, the problem of pathological blood cell classification was addressed through the use of novel deep-learning strategies. We curated a data set for RBC Morphology classification, consisting of samples from three different sources. The process involved preprocessing techniques to establish a data set aligned with our research objectives; 2 other data sets were acquired, targeted for Lymphoma and Leukemia classifications separately.
Three distinct individual models were applied for each of the data sets: the EfficientNetB3, ResNet50V2, and a pre-trained ViT model. To leverage the strengths of both the CNN and ViT architectures, an Ensemble model using the weighted average method was developed.
The present findings confirm that the proposed HematoFusion model mitigates the shortcomings of the individual models by enhancing the accuracy, precision, and sensitivity, achieving more consistent results across the three data sets. While HematoFusion demonstrates competitive or superior performance on Leukemia and Lymphoma classification, particularly in precision and F1-score, it performs

comparably on RBC classification, despite its higher number of classes and the issue of data imbalance that resulted in a few cases of overfitting. We additionally acknowledge certain limitations in predicting a couple of classes. These are the key components to overcome in future research. Future studies should also be devoted to covering more pathological blood disorders and implementing further processing and data augmentations to alleviate the issue of class imbalance and overfitting.

Overall, this paper provides a foundation for future developments by establishing baseline data that future researchers can expand upon to address the limited data available for RBC Morphology and combining the strengths of the residual networks and vision transformers for a more robust framework.

References

[1] Yann LeCun et al. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324. DOI: https://doi.org/10.1109/5.726791.

[2] KS Kim et al. "Analyzing blood cell image to distinguish its abnormalities". In: *Proceedings of the eighth ACM international conference on multimedia*. New York: Association for Computing Machinery, 2000, pp. 395–397. DOI: https://doi.org/10.1145/354384.354543.

[3] Jia Deng et al. "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Miami: Ieee, 2009, pp. 248–255. DOI: https://doi.org/10.1109/CVPR.2009.5206848.

[4] Nikita Orlov et al. "Automatic Classification of Lymphoma Images With Transform-Based Global Features". In: *IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society* 14 (2010), pp. 1003–13. DOI: https://doi.org/10.1109/TITB.2010.2050695.

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira et al. Vol. 25. Lake Tahoe, Nevada: Curran Associates, Inc., 2012, pp. 1097–1105.

[6] K. He et al. *Deep Residual Learning for Image Recognition*. Preprint at https://arxiv.org/abs/1512.03385. 2015.

[7] Mahsa Lotfi et al. "The detection of dacrocyte, schistocyte and elliptocyte cells in iron deficiency anemia". In: *2015 2nd International conference on pattern recognition and image analysis (IPRIA)*. Rasht, Iran: IEEE, 2015, pp. 1–5. DOI: https://doi.org/10.1109/PRIA.2015.7161628.

[8] O. Russakovsky et al. *ImageNet Large Scale Visual Recognition Challenge*. Preprint at https://arxiv.org/abs/1409.0575. 2015.

[9] J. C. Chapin and M. T. Desancho. "Hematologic Dysfunction in the ICU". In: *Critical Care*. Ed. by J. M. Oropello, S. M. Pastores, and V. Kvetan. New York: McGraw-Hill Education, 2016.

[10] Kaiming He et al. *Identity Mappings in Deep Residual Networks*. Preprint at http://arxiv.org/abs/1603.05027. 2016.

[11] Kenneth Kaushansky et al. *Williams Hematology*. New York: McGraw-Hill Education, 2016.

[12] Gao Huang et al. "Densely Connected Convolutional Networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Honolulu: IEEE, 2017, pp. 4700–4708. DOI: https://doi.org/10.48550/arXiv.1608.06993.

[13] Yiyue Jiang et al. "Label-free detection of aggregated platelets in blood by machine-learning-aided optofluidic time-stretch microscopy". In: *Lab on a Chip* 17.14 (2017), pp. 2426–2434. DOI: https://doi.org/10.1039/C7LC00396J.

[14] Mazin Z Othman, Thabit S Mohammed, and Alaa B Ali. "Neural network classification of white blood cell using microscopic images". In: *International Journal of Advanced Computer Science and Applications* 8.5 (2017), pp. 99–103. DOI: https://doi.org/10.14569/IJACSA.2017.080513.

[15] Mohammad Fadly Syahputra, Anita Ratna Sari, and Romi Fadillah Rahmat. "Abnormality classification on the shape of red blood cells using radial basis function network". In: *2017 4th International Conference on Computer Applications and Information Processing Technology (CAIPT)*. Kuta Bali, Indonesia: IEEE, 2017, pp. 1–5. DOI: https://doi.org/10.1109/CAIPT.2017.8320739.

[16] Ashish Vaswani et al. "Attention is all you need". In: *Advances in Neural Information Processing Systems* 30 (2017). DOI: https://doi.org/10.48550/arXiv.1706.03762.

[17] Hajara Abdulkarim Aliyu et al. "Red blood cell classification: deep learning architecture versus support vector machine". In: *2018 2nd international conference on biosignal analysis, processing and systems (ICBAPS)*. Kuching, Malaysia: IEEE, 2018, pp. 142–147. DOI: https://doi.org/10.1109/ICBAPS.2018.8527398.

[18] Paul Mooney. *Blood Cell Images*. 2018. URL: https://www.kaggle.com/datasets/paultimothymooney/blood-cells.

[19] Mariam Nassar et al. "Label-free identification of white blood cells using machine learning". In: *Cytometry Part A* 95.8 (2019), pp. 836–842. DOI: https://doi.org/10.1002/cyto.a.23794.

[20] N. C. Shenggan. *BCCD Dataset*. https://github.com/Shenggan/BCCD_Dataset. 2019.

[21] Mingxing Tan and Quoc V. Le. "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks". In: *Proceedings of the 36th International Conference on Machine Learning*. Vol. 97. Long Beach, California: PMLR, 2019, pp. 6105–6114. DOI: https://doi.org/10.48550/arXiv.1905.11946.

[22] Laith Alzubaidi et al. "Classification of red blood cells in sickle cell anemia using deep convolutional neural network". In: *Intelligent Systems Design and Applications*. Ed. by Ajith Abraham et al. Vol. 1. Cham: Springer International Publishing, 2020, pp. 6–8. DOI: https://doi.org/10.1007/978-3-030-16657-1_51.

[23] Yasmin M Kassim et al. "Clustering-Based Dual Deep Learning Architecture for Detecting Red Blood Cells in Malaria Diagnostic Smears". In: *IEEE Journal of Biomedical and Health Informatics* 25.5 (2020), pp. 1735–1746. DOI: https://doi.org/10.1109/JBHI.2020.3034863.

[24] Tatiana Cristina Figueira Polo and Hélio Amante Miot. *Use of ROC curves in clinical and experimental studies*. 2020. DOI: https://doi.org/10.1590/1677-5449.200186.

[25] Thomas Wolf et al. *HuggingFace's Transformers: State-of-the-art Natural Language Processing*. 2020. arXiv: 1910.03771 [cs.CL]. URL: https://arxiv.org/abs/1910.03771.

[26] Yassine Barhoumi and Ghulam Rasool. *Scopeformer: n-CNN-ViT hybrid model for intracranial hemorrhage classification*. 2021. DOI: https://doi.org/10.48550/arXiv.2107.04575.

[27] Ferhat Bozkurt. "Classification of blood cells from blood cell images using dense convolutional network". In: *Journal of Science, Technology and Engineering Research* 2.2 (2021), pp. 81–88. DOI: https://doi.org/10.53525/jster.1014186.

[28] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. Preprint at https://arxiv.org/abs/2010.11929. 2021.

[29] Mawaddah Harahap et al. "Implementation of Convolutional Neural Network in the classification of red blood cells have affected of malaria". In: *Sinkron: jurnal dan penelitian teknik informatika* 5.2 (2021), pp. 199–207. DOI: https://doi.org/10.33395/sinkron.v5i2.10713.

[30] Danish Jamil et al. "Diagnosis of gastric cancer using machine learning techniques in healthcare sector: a survey". In: *Informatica* 45.7 (2021). DOI: https://doi.org/10.31449/inf.v45i7.3633.

[31] JGraph. *diagrams.net, draw.io*. Oct. 2021. URL: https://www.diagrams.net/.

[32] Zhencun Jiang et al. "Method for diagnosis of acute lymphoblastic leukemia based on ViT-CNN ensemble model". In: *Computational Intelligence and Neuroscience* 2021.1 (2021), p. 7529893. DOI: https://doi.org/10.1155/2021/7529893.

[33] Korranat Naruenatthanaset et al. *Red Blood Cell Segmentation with Overlapping Cell Separation and Classification on Imbalanced Dataset*. Preprint at https://arxiv.org/abs/2012.01321. 2021.

[34] Maithra Raghu et al. "Do vision transformers see like convolutional neural networks?" In: *Advances in neural information processing systems* 34 (2021), pp. 12116–12128. DOI: https://doi.org/10.48550/arXiv.2108.08810.

[35] Georg Steinbuss et al. "Deep learning for the classification of non-Hodgkin lymphoma on histopathological images". In: *Cancers* 13.10 (2021), p. 2419. DOI: https://doi.org/10.3390/cancers13102419.

[36] Željko Vujović et al. "Classification model evaluation metrics". In: *International Journal of Advanced Computer Science and Applications* 12.6 (2021), pp. 599–606. DOI: https://doi.org/10.14569/IJACSA.2021.0120670.

[37] Xufeng Yao et al. "Classification of white blood cells using weighted optimized deformable convolutional neural networks". In: *Artificial Cells, Nanomedicine, and Biotechnology* 49.1 (2021), pp. 147–155. DOI: https://doi.org/10.1080/21691401.2021.1879823.

[38] Kai Jiang et al. "The encoding method of position embeddings in vision transformer". In: *Journal of Visual Communication and Image Representation* 89 (2022), p. 103664. DOI: https://doi.org/10.1016/j.jvcir.2022.103664.

[39] Zahra Mousavi Kouzehkanan et al. "A large dataset of white blood cells containing cell locations and types, along with segmented nuclei and cytoplasm". In: *Scientific Reports* 12.1 (2022), p. 1123. DOI: https://doi.org/10.1038/s41598-021-04426-x.

[40] Dyah Aruming Tyas et al. "Erythrocyte (red blood cell) dataset in thalassemia case". In: *Data in Brief* 41 (2022), p. 107886. DOI: https://doi.org/10.1016/j.dib.2022.107886.

[41] Mohammed Hamdi et al. "Hybrid Models Based on Fusion Features of a CNN and Handcrafted Features for Accurate Histopathological Image Analysis for Diagnosing Malignant Lymphomas". In: *Diagnostics* 13.13 (2023), p. 2258. DOI: https://doi.org/10.3390/diagnostics13132258.

[42] Rojina Kashefi et al. *Explainability of Vision Transformers: A Comprehensive Review and New Perspectives*. Preprint at https://arxiv.org/abs/2311.06786. 2023.

[43] Gireen Naidu, Tranos Zuva, and Elias Mmbongeni Sibanda. "A review of evaluation metrics in machine learning algorithms". In: *Computer science on-line conference*. Springer. 2023, pp. 15–25. DOI: https://doi.org/10.1007/978-3-031-35314-7_2.

[44] Austin H Routt et al. "Deep ensemble learning enables highly accurate classification of stored red blood cell morphology". In: *Scientific Reports* 13.1 (2023), p. 3152. DOI: https://doi.org/10.1038/s41598-023-30214-w.

[45] Hongwei Shang et al. "Precision/recall on imbalanced test data". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2023, pp. 9879–9891. URL: https://proceedings.mlr.press/v206/shang23a.html.

[46] Enquan Yang et al. "DRNet: Dual-stage refinement network with boundary inference for RGB-D semantic segmentation of indoor scenes". In: *Engineering Applications of Artificial Intelligence* 125 (2023), p. 106729. ISSN: 0952-1976. DOI: https://doi.org/10.1016/j.engappai.2023.106729.

[47] Areen K Al-Bashir, Ruba E Khnouf, and Lamis R Bany Issa. "Leukemia classification using different CNN-based algorithms-comparative study". In: *Neural Computing and Applications* 36.16 (2024), pp. 9313–9328. DOI: https://doi.org/10.1007/s00521-024-09554-9.

[48] Martin Moller. "Efficient training of feed-forward neural networks". In: *Neural Network Analysis, Architectures and Applications*. CRC Press, 2024, pp. 136–173. DOI: https://doi.org/10.1201/9781003572886-8.

[49] Emine Özgür and Ahmet Saygılı. "A new approach for automatic classification of non-hodgkin lymphoma using deep learning and classical learning methods on histopathological images". In: *Neural Computing and Applications* 36.32 (2024), pp. 20537–20560. DOI: https://doi.org/10.1007/s00521-024-10229-8.

[50] Sajida Perveen et al. "A framework for early detection of acute lymphoblastic leukemia and its subtypes from peripheral blood smear images using deep ensemble learning technique". In: *IEEE Access* 12 (2024), pp. 29252–29268. DOI: https://doi.org/10.1109/ACCESS.2024.3368031.

[51] Prakeerth Prasad and Jani Anbarasi L. "Acute Lymphoblastic Leukemia Subtypes Detection using Vision Transformer Model". In: *2024 5th International Conference on Data Intelligence and Cognitive Informatics (ICDICI)*. 2024, pp. 1413–1418. DOI: https://doi.org/10.1109/ICDICI62993.2024.10810888.

[52] Ruaa Sadoon and Adala Chaid. "Classification of pulmonary diseases using a deep learning stacking ensemble model". In: *Informatica* 48.14 (2024). DOI: https://doi.org/10.31449/inf.v48i14.6145.

[53] Umair Saeed et al. "DeepLeukNet—A CNN based microscopy adaptation model for acute lymphoblastic leukemia classification". In: *Multimedia Tools and Applications* 83.7 (2024), pp. 21019–21043. DOI: https://doi.org/10.1007/s11042-023-16191-2.

[54] Md Shahin Ali et al. "A Hybrid VGG16-ViT Approach With Image Processing Techniques for Improved White Blood Cell Classification and Disease Diagnosis: A Retrospective Study". In: *Health Science Reports* 8.6 (2025), e70859. DOI: https://doi.org/10.1002/hsr2.70859.

[55] Sazzli Kasim et al. "Multiclass leukemia cell classification using hybrid deep learning and machine learning with CNN-based feature extraction". In: *Scientific Reports* 15.1 (2025), p. 23782. DOI: https://doi.org/10.1038/s41598-025-05585-x.

[56] Aniel Mahendren et al. "White Blood Cells Classification: A Feature-Based Transfer Learning Approach". In: *Selected Proceedings from the 2nd International Conference on Intelligent Manufacturing and Robotics, ICIMR 2024, 22-23 August, Suzhou, China*. Ed. by Wei Chen et al. Singapore: Springer Nature Singapore, 2025, pp. 757–763. ISBN: 978-981-96-3949-6. DOI: https://doi.org/10.1007/978-981-96-3949-6_63.

[57] Mouna Saadallah. *Red Blood Cell Morphology Dataset for Image Classification*. Zenodo, Feb. 2025. DOI: https://doi.org/10.5281/14936017. URL: https://zenodo.org/records/14936017.

[58] Vera Sorin et al. "Deep Learning Applications in Lymphoma Imaging". In: *Acta Haematologica* (2025). DOI: https://doi.org/10.1159/000547427.

[59] KP Swain, SK Swain, and SR Nayak. "Vision Transformer-Based Automated Classification of Acute Lymphoblastic Leukemia". In: *2025 International Conference on Emerging Systems and Intelligent Computing (ESIC)*. IEEE. 2025, pp. 584–588. DOI: https://doi.org/10.1109/ESIC64052.2025.10962707.

[60] Vishesh Tanwar et al. "Enhancing blood cell diagnosis using hybrid residual and dual block transformer network". In: *Bioengineering* 12.2 (2025), p. 98. DOI: https://doi.org/10.3390/bioengineering12020098.