IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 16, 2020, accepted May 24, 2020, date of publication June 2, 2020, date of current version June 11, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2999349

Synthetic Blood Smears Generation Using Locality Sensitive Hashing and Deep Neural Networks

RABIAH AL-QUDAH AND CHING Y. SUEN, (Life Fellow, IEEE)
Department of Computer Science, Concordia University, Montréal, QC H3G 1M8, Canada
Corresponding author: Rabiah Al-Qudah (r_alquda@encs.concordia.ca)

ABSTRACT Peripheral Blood Smear (PBS) analysis is a vital routine test carried out by hematologists to assess some aspects of humans' health status. PBS analysis is prone to human errors and utilizing computer-based analysis can greatly enhance this process in terms of accuracy and cost. Recent approaches in learning algorithms, such as deep learning, are data hungry, but due to the scarcity of labeled medical images, researchers had to find viable alternative solutions to increase the size of available datasets. Synthetic datasets provide a promising solution to data scarcity, however, the complexity of blood smears' natural structure adds an extra layer of challenge to its synthesizing process. In this work, we propose a methodology that utilizes Locality Sensitive Hashing (LSH) to create a novel balanced dataset of 2500 synthetic blood smears. This dataset, which was automatically annotated during the generation phase, will be made public for research purposes and covers 17 essential categories of blood cells. We proved the effectiveness of the proposed dataset by utilizing it for training a deep neural network, this model got a very high accuracy score of 98.72% when tested with the well known ALL-IDB dataset. The dataset also got the approval of 5 experienced hematologists to meet the general standards of making thin blood smears.

INDEX TERMS Automatic annotation, blood films, blood smears, deep learning, LSH, medical data, synthetic dataset.

I. INTRODUCTION

A Blood test is an examination of a sample of blood performed in a specialized medical laboratory by specialists. Blood samples are vital sources of information for illness diagnosis, drug detection, and measurement in human bodies.

There are three main types of blood cells: Red Blood Cells (RBCs), White Blood Cells (WBCs), and Platelets. RBCs, also known as erythrocytes, are in charge of carrying oxygen and carbon dioxide to the entire body. WBCs, also known as leukocytes, are the primary defense system against infectious diseases. Platelets, also known as thrombocytes, are non-nucleated entities responsible for repairing blood vessels in case of injury. RBCs are the most plentiful type of blood cells, For instance, the number of WBCs in adult males ranges from 4.5 to 11.5 thousand in 1 microliter, where the number of RBCs in adult males ranges from 4.6 to 6 million in 1 microlitre [1].

Most blood tests are conducted by only placing a tube filled with liquid blood sample in an automatic analyzer that produces highly accurate results. In some cases, this procedure is not reliable enough and a blood smear is needed to further analyze the sample. A Peripheral Blood Smear (PBS), also known as a blood film, is the result of spreading and staining a thin layer of blood on a glass microscope slide. PBSs are used for three main purposes:

1) Verify automated analyzer results.
2) Identify atypical, immature, and abnormal cells.
3) Identify morphological abnormalities that are beyond the capabilities of the automated analyzers.

The automation of blood smear analysis has attracted the attention of researchers in recent years, As the automation of this procedure can save medical specialists time and lab consumables. This is more apparent at epidemics and pandemics times, as saving time and consumables are more crucial. The automation can also aid in delivering more accurate results, as the recognition of some morphological abnormalities is challenging even for experts. And despite the good results achieved so far in this context, many challenges still arise. In this work we tackle a main obstacle faced by computer researchers who work on automating blood analyses. The main contributions of this work are:

The associate editor coordinating the review of this manuscript and approving it for publication was Kemal Polat.

1) Employing Locality Sensitive Hashing (LSH) with Random Projections as a synthetic image generation method.
2) Creating a dataset of 2500 synthetic whole blood smears, that that will be made public for research purposes, as it is not restricted by any privacy constraints. To our knowledge this is the first comprehensive synthetic dataset of this kind. Beside getting the approval of five medical experts for this dataset, we also prove its effectiveness as a training set for classification networks.
3) Providing two sets of annotations for the proposed dataset, that were automatically generated while constructing the smear images.

II. MOTIVATION

Deep Neural Networks are reliant on large volumes of data. Obtaining large datasets can be a real challenge for researchers from different research areas. Acquiring medical datasets can be even more challenging due to privacy constraints on patients' data. Moreover, annotating this type of data is a costly procedure that can only be performed by medical experts. In the context of blood smear analysis some extra domain-specific challenges arise. The first challenge is that some blood cell subtypes are rare in occurrence, for example, the authors in [2], were able to collect only three samples of reactive plasmacytosis in three years. Hence, having a sufficient number of blood smears containing such rare types for training a deep network might take many years. Additionally, the complexity of preparing a balanced dataset of blood smears comes from its natural structure; each blood smear contains hundreds of blood cells from different types, which are naturally distributed in an imbalanced manner. In [3], the authors demonstrated that RBCs occurred approximately seven times more than WBCs in the training set. This challenge implies that traditional augmentation techniques might not help as it will only amplify the imbalance issue. Figure 1 shows the imbalance between the main blood cell types, the few WBCs cells are colored in purple, whilst the other cells are of type RBCs.

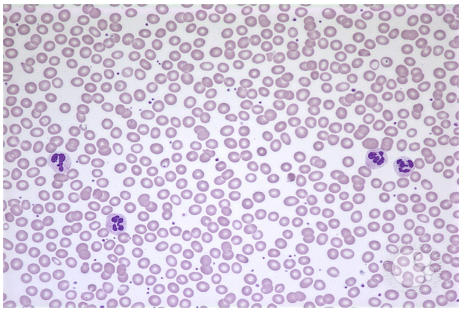

FIGURE 1. Whole-slide images [5].

Moreover, most datasets employed in this research area are private which limits results reproducibility and comparability. Besides, most public blood smear datasets are for segmented

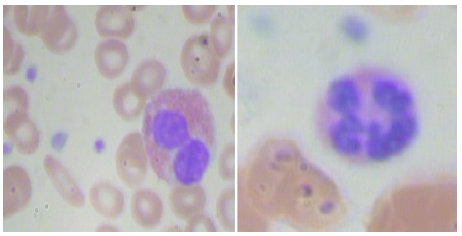

FIGURE 2. Segmented microscopic images [6].

blood cells, see Figure 2, this means that only one blood cell appears in each instance, which causes inconvenience in real applications [4].

Finally, the available public datasets are only annotated for the main blood cell types; RBCs, WBCs, and Platelets or the main normal types of WBCs which is not sufficient to fully comprehend and analyse blood smears. These challenges usually leave researchers with two choices, either narrow down research scope or go through the costly process of making and annotating a new dataset which might not have sufficient instances from all types (i.e., imbalanced dataset).

Generating synthetic blood smears can provide a solution to the challenges mentioned above, as it is not constrained by any privacy issues and it is possible to control the number of labelled instances of each cell type.

TABLE 1. Types of blood cells.

Main blood cell type	Blood cell subtypes
Platelets	Giant Platelets, Activated Platelets
WBCs	Atypical Lymphocyte, Atypical Monocyte, Band cell, Basket cell, Basophil, Eosinophil, Hyper-Segmented Neutrophils, Hypo-Segmented Neutrophils, Lymphocyte, Monocyte, Plasma cell, Reactive Lymphocyte, Segmented Neutrophils
RBCs	Nucleated RBCs

As mentioned in Section I, the main purpose of blood smears is studying abnormal cells and morphologies, rather than normal and main blood cell types. Table 1 summarises each main blood cell type and its corresponding subtypes. The target of PBS analysis is to classify the categories in the blood cell subtypes column. This has been automated in the literature either by one step approaches [7] or multi-step pipelines [3]. Hence, in this study both approaches will be taken into consideration, and two sets of annotations will be automatically composed during image constructions:

1) The first set of annotations targets multi-step pipelines, which classify the main types of blood cells, then further classify the regions of interest into its corresponding subtypes. Since Nucleated RBCs is the only subcategory of RBCs, this annotation set will classify cells into: WBCs, Platelets, and Nucleated RBCs.
2) The second set of annotations targets one-step classifiers, where the system classifies whole blood smears into all 16 subtypes. Platelets category is also annotated

in this set based on the advice of a medical expert. Hence, this set of annotations includes 17 categories.

Throughout this study we will refer to WBCs, Nucleated RBCs, and Platelets as the main blood cell types, and will refer to the rest of the subtypes and abnormal morphologies as blood cell subtypes as listed in Table 1. Next, we discuss viable approaches towards constructing and engineering a synthetic blood smear dataset.

III. RELATED WORK

Blood smear analysis has been an active research topic, that has attracted the attention of medical experts [8], [9] and computer scientists [10], [11] over the years.

Automatic PBS analysis has been utilized as a Leukemia diagnosis tool in [12] with 100% accuracy, where three deep architectures (AlexNet, CaffeNet, Vgg-f) were trained to generate features from PBS images. All features were concatenated and reduced by applying the gain ratio algorithm, before being emitted to a Support Vector Machine (SVM) classifier. The authers in [13] applied many augmentation operations on the ALL-IDB public dataset [14], such as, histogram equalization, translation, reflection, rotation, shearing, conversion to grayscale, and blurring. A convolutional neural network of 5 convolutional layers, a fully connected layer and a softmax layer was trained with the augmented dataset and achieved an accuracy score of 96.6%.

Automatic PBS analysis has also been applied to Malaria detection. The authors in [15] developed an Android smartphone application using a dataset of whole slide thick smear images. The methodology proposed in this work reduces the size of the initial search space by applying an intensity-based Iterative Global Minimum Screening (IGMS) procedure, all regions of interest are then directed to a CNN model consisting of seven convolutional layers. The classification accuracy was 93.46%. The authors in [16] utilized a dataset of whole slide peripheral blood smears to train a Deep Belief Network (DBN). A concatenated feature of color and texture was used to initialize the visible layer of the 4 hidden layer DBN. The network achieved an F-score of 89.66.

In some works automated PBS analysis was presented as a tool for counting and classifying blood cells. The work in [17] classified the five main normal types of WBCs in medical hyperspectral imaging (MHSI). Four different architectures were utilized for this purpose: SVM, VGG16, CNN without Gabor wavelet, CNN with Gabor wavelet and a combination of modulated Gabor wavelet and CNN kernels, named as MGCNN. The proposed model achieved its highest accuracy score of 97.65%. The work in [18] combined Fourier Ptychographic Microscopy (FPM) and an adjusted version of You Only Look Once (YOLO) network for the purpose of WBC detection. In the proposed YOLO network, the feature maps of the last three layers were concatenated and passed to a final convolution layer.

Other work in the literature, tackled a deeper level of cell classification; the work in [7] for example, classified 40 types and abnormal morphologies of blood by training a residual deep network. The average classification accuracy was 76.84%.

Recent approaches in learning algorithms like deep learning are data hungry, but due to the scarcity of labeled medical images [19], researchers had to find viable solutions to increase the size of available datasets like augmentation [20], [21]. A shortcoming of these augmentation techniques is that it is performed by trial and error, and there is no guarantee that it will enhance results until after training, which might lead to repeating the training process [22]. An alternative approach can be creating realistic instances, i.e. synthetic instances. The work in [23] presented the first synthetic blood smears dataset which was created considering only RBCs. This work presents a promising solution for the problem of data scarcity, and despite its importance, it is not sufficiently considered in the literature and there is still room for improvement in this context.

IV. METHODOLOGY

In this section, we propose a framework for constructing a synthetic balanced dataset of blood smears. Synthetic data have helped to improve the performance of Neural Networks (NN) by providing sufficient instances that help NNs learn features of target classes. Synthetic instances do not necessarily need to look identical to real instances, but it must look realistic, hence, the quality of the smears produced by our framework does not matter as much as its ability to help the classification network better generalize [22].

The proposed framework aims to assemble images of segmented blood cells from all main and sub cell types as mentioned in Table 1 on blood smear canvases while keeping in mind the dataset balancing issue and retaining the natural distribution of blood cells. This approach, as illustrated in Figure 3, runs in two phases; Data pools preparation phase, and blood smears generation and annotation phase. The expected results of the proposed approach are:

1) A dataset of blood smear images.
2) A set of annotation files for the datastet instances that annotates blood cells to three main classes (Nucleated RBC, WBC, Platelet)
3) A set of annotation files for the datastet instances that annotates blood cells to the 17 subtypes of blood cells.

The following subsections provide more details about each phase.

A. PHASE 1: DATA POOLS PREPARATION

In this phase, we aim to create an image pool for each main and sub blood cell type, hence 18 image pools will be processed and ready by the end of this phase; 17 blood cell subtypes and RBCs. The RBCs class is added as a pool because it still shows in all blood smears even if it is not annotated. Each of the 18 pools contains images of segmented cells of the pool cell type. All images were collected from public resources and processed by removing the background. All images were

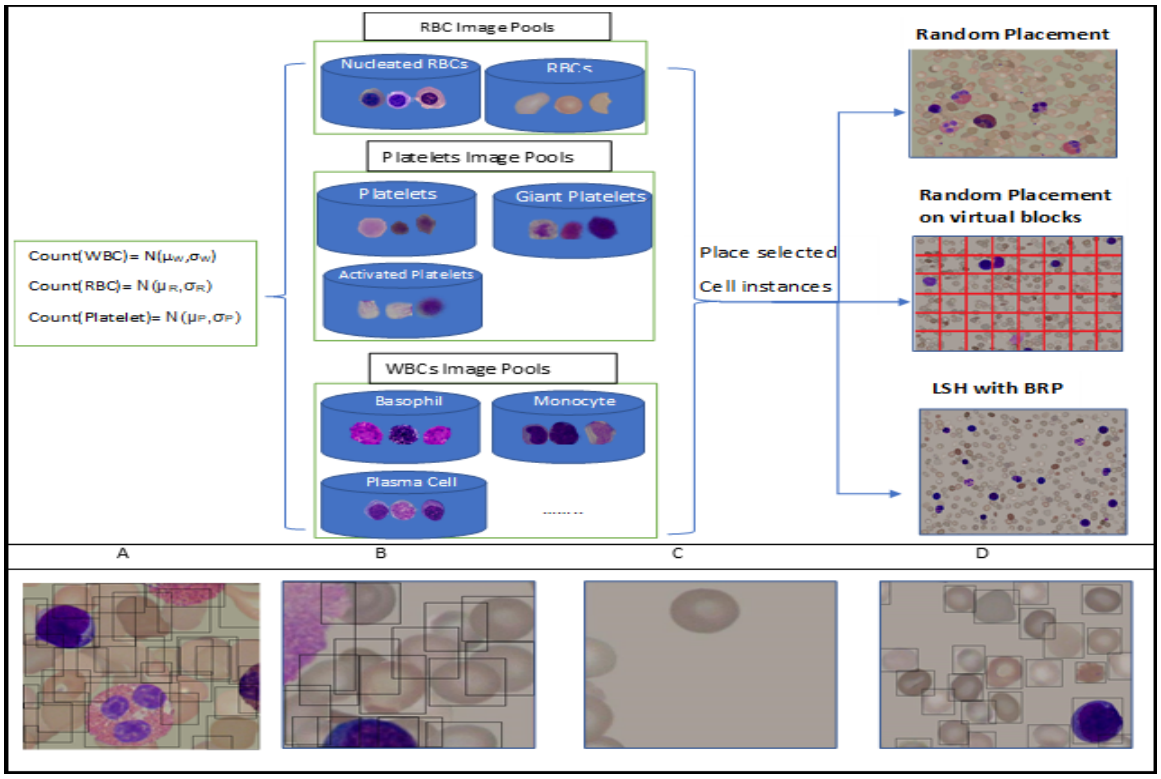

FIGURE 3. Constructing Blood Smears approaches as Black Box. *A* is an enlarged annotated sample result from the approach described in IV-B1, *B* and *C* are enlarged sample results from the approach described in IV-B2, *D* is an enlarged sample result from the approach described IV-B3.

augmented by multiple rotations. Some restrictions need to be considered at this stage as follows:

1) All images must be subject to the same microscope magnification, because the size of the blood cell can be a major factor in distinguishing and classifying some morphological abnormalities.
2) All instances must be treated with the same stain for consistency purposes.
3) Each pool must be representative, well generalised, and comprise all possible appearances of the cell type. In other words, for each cell type we aim to collect distinct images that cover all possible features more than we aim to have a large number of instances.

B. PHASE 2: BLOOD SMEARS GENERATION AND ANNOTATION

In this phase, cell images from phase one are assembled on blood smear canvases to form thin blood smear instances. This procedure can be a bit complicated because each blood smear contains hundreds or cells that have to be ordered in a natural realistic way. The following factors need to be considered at this phase:

1) No restrictions limit the selection of blood cell subtypes to appear in a blood smear, as the presence, absence or deficiency of each cell type or subtype represents certain types of syndromes or medical diagnoses that is independent from all other syndromes or medical diagnoses that can be concluded from other cell types appear in the same blood smear.
2) Total number of the main blood cell types must be carefully selected to represent realistic blood smears. Hence, RBCs, Platelets and WBCs cells are assumed to follow Normal (Gaussian) Distributions. The Normal distributions of WBCs and Platelets and their parameters were derived from the All-IDB dataset [14] statistics, However, the RBCs' Gaussian Distribution was assumed to follow the ones as in [23].

The Gaussian Distribution is a probability distribution that typically used to model normal phenomena and is described by the probability density function (PDF). A PDF describes the probability of a value (x) of an experiment to fall within a particular range of values, it is mathematically represented by the following formula:

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \tag{1}$$

where, the mean denoted by μ is the Arithmetic average of data, and σ is the standard deviation that is calculated by:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{N} \tag{2}$$

TABLE 2. Statistics of the main blood cell types.

Blood cell type	Mean	Standard Deviation
Platelets	4.3	4.8
WBCs	8.6	10.5
RBCs	669	149

where, N is the number of sample observations. The bell curve of a Gaussian Distribution is centered and symmetric around the mean and stretched by the standard deviation. Table 1 lists μ and σ for each main cell type.

At this point an efficient approach is needed to place cell's images from different pools on a blood smear background canvas. For this purpose, we first select three random numbers; a random number from each Gaussian Distribution. Second, subtypes of each main blood cell type are uniformly selected. Next, instances from each selected subtype pool are selected. When all subtypes and cell instances are selected from image pools, an efficient approach is needed to place these cells on canvas. In the following subsections we present and discuss three different placement strategies to paste the selected instances on blood smear canvas in a realistic fashion.

1) NAIVE APPROACH 1: RANDOM PLACEMENT

In this approach random paste coordinates are selected for each cell. This approach executes fast and is easy to implement but it does not guarantee the spread of cells on the smear canvas, instead, as shown in Figure 3-A it forms cell clumps, and some cells will override others which leads to wrong annotations. Figure 3-A shows how bounding boxes are very intersected.

2) NAIVE APPROACH 2: RANDOM PLACEMENT ON VIRTUAL BLOCKS

In this approach the blood smear background canvas is divided into MxN virtual blocks. For each cell a block is uniformly chosen, then random paste coordinates are selected inside the chosen virtual block. a drawback of this approach is that some blocks might be selected more than others which will also form cell clumps. Figure 3-B illustrates a crowded block where 3-C illustrates an almost empty block.

3) NEAREST NEIGHBOUR MINING APPROACH

The main shortcoming resulted from the previous naive approaches is the formation of clumps in the blood smears. To avoid placing cells on canvas with high probability of occlusion we need to choose a paste location that does not overlap with any other surrounding cell. Implementing this in a brute-force manner can guarantee the accuracy of the results but the processing time grows linearly with the number of cells on canvas. On the other hand, choosing a potential paste point and estimating its nearest neighbors and reject those points that will cause occlusion with neighboring objects can reduce the number of comparisons and the complexity. Locality Sensitive Hashing (LSH) is a nearest neighbour retrieval algorithm that can be utilised for this purpose because it is a generic hashing technique that intends to preserve the local relations of the data.

In our problem, we have a set of cell objects to be pasted on a canvas, each object will be represented by its top left coordinates, called paste points, hence our space is a 2D Euclidean space. A dictionary keeps the width and height of the paste point's corresponding cell. Our goal is to retrieve the nearest neighbor points to each potential paste point, check if the dimensions of the paste point will overlap with its neighbors or not within a certain threshold, and finally decide to accept this new point or reject it. An effective approach to implement our goal on the mentioned Euclidean space can be dividing the space by a set of projections and hash near points into the same bucket. This approach is called Random Projections.

The core idea behind random projections is given in the Johnson-Lindenstrauss lemma, [24] which states that if points in a vector space are of sufficiently high dimension, then they may be projected into a suitable lower-dimensional space in a way which approximately preserves the distances between the points. This can be represented as:

Let $\mathbb{X}$ be a space of objects [25], to which dataset and query objects belong. Let D be a distance measure defined on $\mathbb{X}$. Let H be a family of hash functions h: $\mathbb{X} \rightarrow \mathbb{Z}$, where $\mathbb{Z}$ is the set of integers.

Let R_1, R_2, P_1, and P_2 be real numbers. For any points X_1 and X_2 in $\mathbb{X}$ that are close to each other, there is a high probability P_1 that they fall into the same bucket

$$P_H[h(X_1) = h(X_2)] \geq P_1 for D(X_1, X_2) \leq R_1 \tag{3}$$

Moreover, for any points X_1 and X_2 in $\mathbb{X}$ that are far apart, there is a low probability $P_2 < P_1$ that they fall into the same bucket

$$P_H[h(X_1) = h(X_2)] \leq P_2 for D(X_1, X_2) \geq cR_1 = R_2 \tag{4}$$

Let L denote the number of random projections, then the space will be partitioned using L hyperplanes by selecting $pr_1, \ldots, pr_L$ vectors at random from a Gaussian distribution. Then each dataset and query objects are hashed using equation 5, in our work we opted to choose random binary projections.

$$[h(m)]_i = \begin{cases} 0 & pr^T m \leq 0 \\ 1 & pr^T m > 0, \end{cases} \quad i = 1 \ldots L \tag{5}$$

Object m will then be stored in a hash table with its hash value, h(m), as its key. Every time a potential paste point is queried against the LSH engine, it will be hashed using equation 5, and all previous points in the same bucket will be returned as possible neighbors. Each of the returned neighbors will be checked against a distance criteria, if all neighbor points are farther than a certain threshold, then the potential paste point will pass and the object will be pasted. Else, if at least one neighbor point is closer than the same

threshold, then the point will be rejected and a new paste point will be selected and queried against the LSH engine. Since we're dealing with cells as bounding boxes, a good distance measure will be measuring the intersection over union ratio, a.k.a Jaccard similarity between the potential paste point and its neighbors. Jaccard Similarity can be defined as

$$J(p, n_i) = |p \cap n_i| / |p \cup n_i| \tag{6}$$

where n_i denotes a neighbor point, and p denotes the potential paste point. The Jaccard similarity value of between each neighbor and the paste point will be checked by equation 7 to check the validity of the paste point.

$$Evaluation[p] = \begin{cases} 0 & \exists n_i, \quad J(p, n_i) > T \\ 1 & Else, \end{cases} \qquad n_i \in N \tag{7}$$

Algorithm 1 demonstrates more details of the proposed approach. The following parameters are defined in Algorithm 1:

- R_w, R_R, R_P are the random counts of WBCs, RBCs, and Platelets respectively. Each of these counts represents the number of cells that will appear in the being generated blood smear instance.
- *LSH*: is the locality sensitive hashing engine, that is initialized once instantiated to D dimensions and L projections.
- *Max*: for each of the main cell types (WBCs, Platelets, Nucleated RBCs), its corresponding subtypes have to appear in percentages that sum up to 1. The *Max* parameter is initialized to 1, and is later decreased as the algorithm progresses.

Some highlights from Algorithm 1 are:

- In lines 6 to 9: for each of the main blood cell types, subtypes are randomly chosen and stored in the *SubTypesCount* parameter. These subtypes and the percentages of which each of them will contribute in the blood smear instance are then randomly selected.
- In lines 11 to 15: cell instances are selected from corresponding pools, one cell instance at a time, and a random paste point is then selected within the blood smear canvas. Next, the LSH engine retrieves all potential neighbor points.
- In lines 16 to 22: each neighbor is checked against the potential paste point using the Jaccard similarity metric. If any neighbor overlaps with the potential point past the allowed percentage then the potential paste point will be rejected.

Algorithm 1 Creating a Blood Smear Dataset of Size DS

```
1:  for i ← 1 to DS do
2:      R_W ← N(σ_W, μ_W)
3:      R_R ← N(σ_R, μ_R)
4:      R_P ← N(σ_P, μ_P)
5:      LSH ← D(Dimensions), L(Projections)
6:      for R ← R_W, R_R, R_P do
            SubTypesCount ← Rand(1, Max_types)
7:          Max ← 1
8:          for S ← 1 to SubTypesCount do
                TypePercentage ← Rand(Min, Max)
9:              Max ← Max − TypePercentage
10:             for Count ← 1 to TypePercentage ∗ R do
11:                 Cell ← Rand(Pool(S))
12:                 RejectPoint ← 1
13:                 while RejectPoint do
14:                     PastePoint ← Rand(Canvas)
15:                     Neighbors ← LSH.NN(PastePoint)
16:                     for N ← Neighbors do
17:                         if Jaccard(N, PastePoint) ≥ T then
18:                             RejectPoint ← 1
19:                             Break;
20:                         else
21:                             RejectPoint ← 0
22:                         end if
23:                     end for
24:                     if RejectPoint = 0 then
25:                         Paste(PastePoint, Cell, Canvas)
26:                         Anotate(PastePoint, MainTypesFile)
27:                         Anotate(PastePoint, SubTypesFile)
28:                     end if
29:                 end while
30:             end for
31:         end for
32:     end for
33:     SaveAnnotation(MainTypesFile_i)
34:     SaveAnnotation(SubTypesFile_i)
35:     SaveImage(Canvas_i)
36: end for
```

V. EXPERIMENTS AND RESULTS

In this section we present dataset generation experiments and results. We also conduct two sets of experiments to prove the effectiveness of this dataset.

A. SYNTHETIC DATASET GENERATION

To conduct this set of experiments, Algorithm 1 was implemented and executed. During the execution, we noted that the initial set of cells were often pasted without any overlapping and without the need to retrieve neighbors due to space availability. Hence, to further decrease the cost of execution, the initial 60 cells were pasted without neighbor retrieval. The threshold 60 was chosen empirically.

To further tune our algorithm, the assignment of the threshold parameter T from equation 7 was adjusted; instead of setting T to a constant value throughout the execution, we opted to assign it to a random value from an acceptable range that reflects the real distribution pattern. With this improvement, the synthetic blood smear instances reflected more realistic scenarios, For example, some RBCs were stacked in many smears forming patterns similar to the well-known Rouleaux formation. The resulting datasets

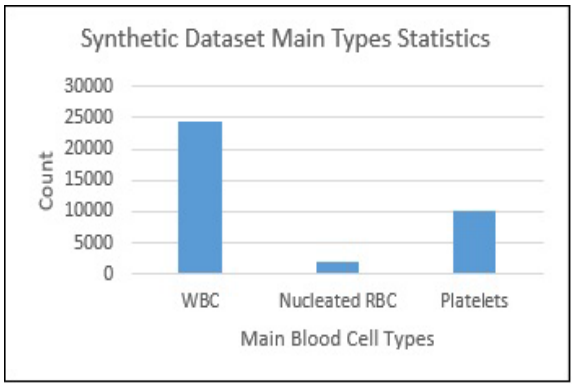

FIGURE 4. Synthetic dataset main types statistics.

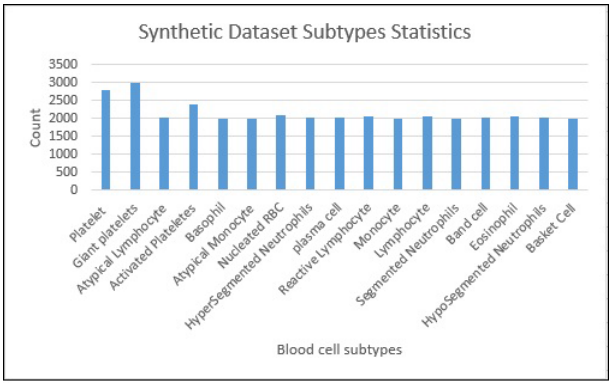

FIGURE 5. Synthetic dataset subtypes statistics.

consists of 2500 blood smears, the dataset is balanced in terms of subtypes and each subtype has around 2000 total instances. Figure 4 and 5 depict more statistical details of the synthetic dataset. Figure 5 demonstrates the dataset balance. Figure 6 shows some samples from the synthetic dataset; smears (a) to (c) belong to the synthetic dataset, while the rest belong to ALL-IDB1 dataset.

B. EXPERIMENTS ON THE SYNTHETIC DATASET

To test the effectiveness of the proposed synthetic dataset we ran two sets of experiments.

1) BLOOD CELL CLASSIFICATION USING DEEP LEARNING

To test the effectiveness of our dataset, we fed the synthetic blood smears along with the first set of annotations; the main types annotation, to three YOLO deep networks. All computations mentioned in this section were made on the supercomputer Helios from Laval University, managed by Calcul Québec and Compute Canada.

YOLO deep network [26], is a real time object detection and classification network that performs objects' detection and classification in one scan.

YOLO is a convolutional neural network which divides the input image into an *NxN* grid. Each grid cell predicts bounding boxes and confidence scores for those boxes, where confidence scores mirror how confident the network is that the bounding box contains an object. Each bounding box consists of 5 attributes: (x, y, h, w, confidence) where, x and y are the coordinates of the center of the bounding box, h, and w are the height and the width of the bounding box. Moreover, Each grid cell predicts C conditional class probabilities to classify the object that is located in the grid cell. A deep network loss function aims to minimize the network error and in YOLO it is calculated as the combination of localisation and classification error:

$$\begin{aligned}
&\sum_{i=0}^{S^2}\sum_{j=0}^{B} \mathbb{1}_{ij}^{obj}[(x_i-\hat{x}_i)^2+(y_i-\hat{y}_i)^2] \\
&+\lambda_{coord}\sum_{i=0}^{S^2}\sum_{j=0}^{B} \mathbb{1}_{ij}^{obj}[(\sqrt{w_i}-\sqrt{\hat{w}_i})^2+(\sqrt{h_i}-\sqrt{\hat{h}_i})^2] \\
&+\lambda_{coord}\sum_{i=0}^{S^2}\sum_{j=0}^{B} \mathbb{1}_{ij}^{noobj}(C_i-\hat{C}_i)^2 \\
&+\lambda_{coord}\sum_{i=0}^{S^2} \mathbb{1}_{i}^{obj}\sum_{c\in classes}(p_i(c)-\hat{p}_i(c))^2 \qquad (8)
\end{aligned}$$

where $\mathbb{1}_{i}^{obj}$ denotes if object appears in cell i and $\mathbb{1}_{ij}^{obj}$ denotes that the jth bounding box predictor in cell i is responsible for that prediction. All experiments were trained with the synthetic dataset and tested with the ALL-IDB1 Dataset. In our first experiment, we trained a Tiny YOLOv3 network [27]. Due to its relatively small size (13 convolutional layers) and the high complexity of the context, the network achieved a low Mean Average Precision (MAP) score of approximately 40%. To improve the results, we trained a larger network, namely YOLOv2 [26] that consists of 23 convolutional layers, This network achieved a better MAP score of 97.59%.

To further improve this result, we randomly resized the network input resolution every 10 batches during training, this regime which was proposed in [26] works like data augmentation and helps the network learn to better generalise. By exposing the network to randomness, the MAP score was improved to 98.72%. Both Nucleated RBCs and WBCs classes were classified with 100% Average Precision (AP), while Platelets scored 96.4%. This network ran with a momentum value of 0.9, learning rate of 0.001, and network input resolution of 800 as an initial resolution. Table 3 summarizes the blood cell classification experiments and results.

2) MEDICAL ASSESSMENT

A questionnaire of 12 questions about the dataset was created by the authors to ensure that the dataset meets the medical standards of thin blood smears. The questionnaire was filled by 5 hematologists with vast years of experience. A blood smear was displayed in each question, followed by some

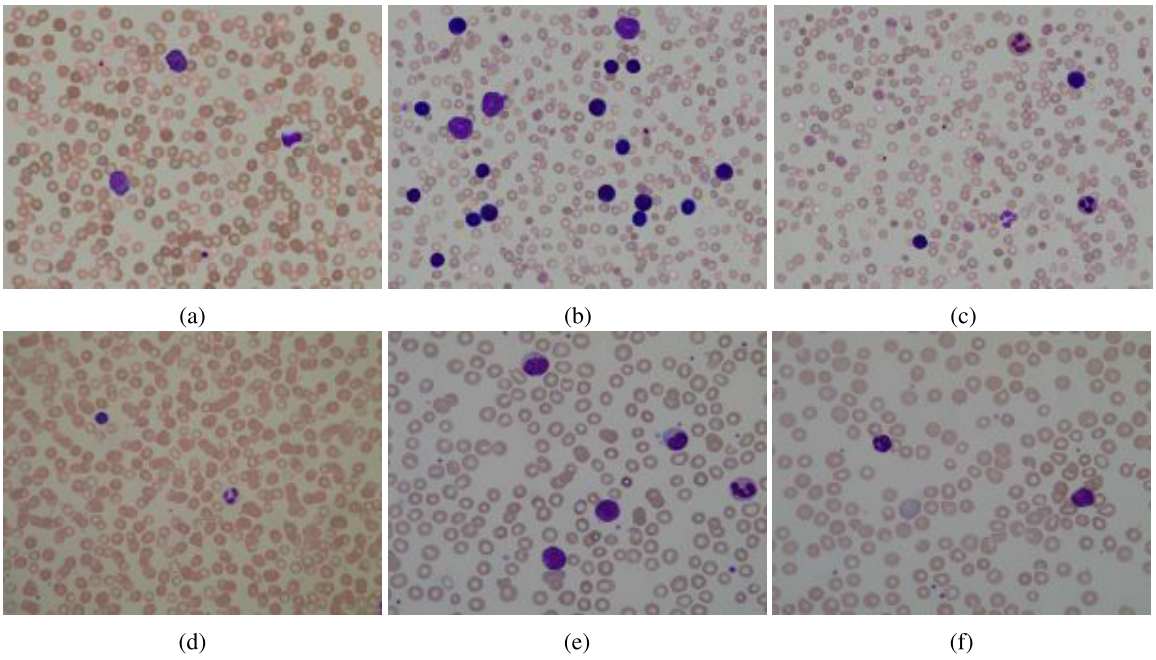

FIGURE 6. Instances from real and synthetic blood smears.

TABLE 3. Blood cell classification results using YOLO.

Deep Network	Mean average precision
Tiny YOLOv3	40%
YOLOv2 without random resizing	97.59%
YOLOv2 with random resizing	98.72%

questions to evaluate the following aspects about the dataset instances:

1) **Factor 1**: The general quality of the smears in terms of the total numbers of RBCs, WBCs and Platelets.
2) **Factor 2**: The correctness of the first set of annotation, by asking the respondents to verify some labels.
3) **Factor 3**: The correctness of the second set of annotation.
4) **Factor 4**: The quality of blood cell subtype choices.
5) **Factor 5**: The level of overlap and occlusion between the blood cells on the synthetic blood smears to assess the quality of the locations produced by our proposed algorithm.

The questionnaire was initially reviewed and verified by a designated hematologist for quality assurance purposes. The average years of experience for the participating hematologists is 6 years. The expected answers were set before starting this experiment and will be denoted by the ground truth throughout this section. Figure 7 demonstrates the results of the questionnaire, where the responses to each question is represented by a column in the chart. In each column, similar answers were grouped by color, and the blue color is used for answers that are identical to the ground truth. Figure 7 shows clearly that all responses were identical and

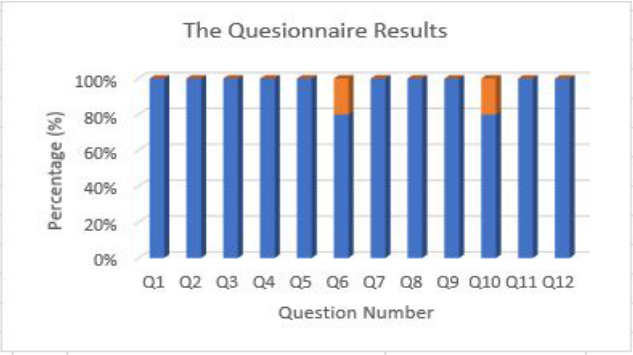

FIGURE 7. The results of the questionnaire.

meet the expected answers in all questions except for question 6 and 10. The sixth question is one of the questions that is used to verify factor 3. It shows a synthetic blood smear and asks the respondent to classify on the platelets that appear in the smear. The platelet is sub-classified as a platelet in the second set of annotations of the proposed dataset, however, one of the hematologists classified it as a giant platelet. It is common that medical experts have different opinions about medical data annotation [28]. The tenth question tests the ratio of the blood cell sizes in the smear. One of the hematologists expressed that some cells appeared a little larger than it should be, where the other 4 hematologists believed that the ratios were appropriate. Factor 5 verifies the correctness of the proposed algorithm. Due to the importance of the last assessment factor, the haematologists were asked to assess all the blood smears that appeared in the questionnaire in terms of the quality of cells distribution and occlusion. All respondents approved that all the slides that were shown

in the questionnaire were thin and all cells were distributed in a natural pattern with a ratio of occlusion that meets the standard one in real blood smears.

The proposed dataset proved its usefulness in the context of PBS analysis, it also provides the following benefits:

1) The dataset is not subject to any privacy or security constraints since it does not belong to real people.
2) Rare subtypes, like Plasma cells, are sufficiently present in the dataset.
3) The dataset is annotated without any extra efforts of medical experts.

VI. CONCLUSION AND FUTURE WORK

In this paper, we provided a novel solution to creating a dataset of synthetic blood smears. Each thin blood smear contains hundreds of blood cells, which leads to a highly complicated synthesizing procedure. This novel dataset consists of 2500 instances, and was automatically annotated for 17 essential blood cell types and abnormal morphologies during the instances generation process. Such synthetic dataset is valuable since labeled medical data is scarce due to extra security and privacy constraints enforced on it. In order to create this dataset, 18 image pools were created in the first phase. In the second phase, RBCs, WBCs and Platelets counts were selected from Gaussian distributions. Next, LSH was employed to divide cells' space into N projections and all near objects were hashed into the same bucket. All cells that hashed in the same bucket were tested against each other using Jaccard similarity, and all cells caused collision higher than an acceptable threshold were rejected. Three YOLO neural networks were trained on the proposed dataset and tested on the ALL-IDB dataset, an accuracy score of 98.72% was achieved. The dataset was also reviewed by a group of highly experienced hematologists from different countries to ensure that it meets the general standards of making thin blood smears.

The Deep network utilised in this work was only trained on the first set of annotations. For future work, we plan to train a Deep classification network on the second set of annotations.

VII. DATASET AVAILABILITY

The proposed synthetic blood smear dataset will be made public for research purposes upon request from the authors.

REFERENCES

[1] F. B. Rodak, A. G. Fritsma, and K. Doig, *Hematology: Clinical Principles and Applications*. Amsterdam, The Netherlands: Elsevier, 2007.
[2] C. Touzeau, C. Pellat-Deceunynck, T. Gastinne, F. Accard, G. Jego, H. Avet-Loiseau, N. Robillard, J. Harousseau, R. Bataille, and P. Moreau, "Reactive plasmacytoses can mimick plasma cell leukemia: Therapeutical implications," *Leukemia Lymphoma*, vol. 48, pp. 207–208, Feb. 2007.
[3] D. Mundhra, B. Cheluvaraju, J. Rampure, and T. R. Dastidar, "Analyzing microscopic images of peripheral blood smear using deep learning," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, M. J. Cardoso, T. Arbel, G. Carneiro, T. Syeda-Mahmood, J. Manuel R. S. Tavares, M. Moradi, A. Bradley, H. Greenspan, J. P. Papa, A. Madabhushi, J. C. Nascimento, J. S. Cardoso, and V. Belagiannis, and Z. Lu, Eds. Cham, Switzerland: Springer, 2017, pp. 178–185.
[4] Q. Wang, S. Bi, M. Sun, Y. Wang, D. Wang, and S. Yang, "Deep learning approach to peripheral leukocyte recognition," *PLoS ONE*, vol. 14, no. 6, Jun. 2019, Art. no. e0218808.
[5] Peter Maslak. *Normal Peripheral Blood Smear*. Accessed: Feb. 24, 2020. [Online]. Available: https://imagebank.hematology.org/image/3666/normal-peripheral-blood-smear–1
[6] Cosmicad. *Bccd Dataset*. Accessed: Feb. 26, 2020. [Online]. Available: https://github.com/cosmicad/dataset
[7] F. Qin, N. Gao, Y. Peng, Z. Wu, S. Shen, and A. Grudtsin, "Fine-grained leukocyte classification with deep residual learning for microscopic images," *Comput. Methods Programs Biomed.*, vol. 162, pp. 243–252, Aug. 2018.
[8] V. B. Shidham and V. K. Swami, "Evaluation of apoptotic leukocytes in peripheral blood smears," *Arch. Pathol. Lab. Med.*, vol. 124, no. 9, pp. 1291–1294, 2000.
[9] V. Moioli, S. Buoro, M. Seghezzi, G. Previtali, P. Dominoni, C. Ottomano, and G. Lippi, "A specific abnormal scattergram of peripheral blood leukocytes suggestive for the presence of proerythroblast," *Scandin. J. Clin. Lab. Invest.*, vol. 80, no. 1, pp. 55–58, Jan. 2020.
[10] S. F. Bikhet, A. M. Darwish, H. A. Tolba, and S. I. Shaheen, "Segmentation and classification of white blood cells," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 4, Jun. 2000, pp. 2259–2261.
[11] C. Di Ruberto, A. Loddo, and L. Putzu, "Detection of red and white blood cells from microscopic blood images using a region proposal approach," *Comput. Biol. Med.*, vol. 116, Jan. 2020, Art. no. 103530.
[12] L. H. S. Vogado, R. M. S. Veras, F. H. D. Araujo, R. R. V. Silva, and K. R. T. Aires, "Leukemia diagnosis in blood slides using transfer learning in CNNs and SVM for classification," *Eng. Appl. Artif. Intell.*, vol. 72, pp. 415–422, Jun. 2018.
[13] T. T. P. Thanh, C. Vununu, S. Atoev, S.-H. Lee, and K.-R. Kwon, "Leukemia blood cell image classification using convolutional neural network," *Int. J. Comput. Theory Eng.*, vol. 10, no. 2, pp. 54–58, 2018.
[14] R. D. Labati, V. Piuri, and F. Scotti, "All-IDB: The acute lymphoblastic leukemia image database for image processing," in *Proc. 18th IEEE Int. Conf. Image Process.*, Brussels, Belgium, Sep. 2011, pp. 2045–2048.
[15] F. Yang, M. Poostchi, H. Yu, Z. Zhou, K. Silamut, J. Yu, R. J. Maude, S. Jaeger, and S. Antani, "Deep learning for smartphone-based malaria parasite detection in thick blood smears," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 5, pp. 1427–1438, May 2020.
[16] D. Bibin, M. S. Nair, and P. Punitha, "Malaria parasite detection from peripheral blood smear images using deep belief networks," *IEEE Access*, vol. 5, pp. 9099–9108, 2017.
[17] Q. Huang, W. Li, B. Zhang, Q. Li, R. Tao, and N. H. Lovell, "Blood cell classification based on hyperspectral imaging with modulated Gabor and CNN," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 1, pp. 160–170, Jan. 2020.
[18] X. Wang, T. Xu, J. Zhang, S. Chen, and Y. Zhang, "SO-YOLO based WBC detection with Fourier ptychographic microscopy," *IEEE Access*, vol. 6, pp. 51566–51576, 2018.
[19] M. Chen, X. Shi, Y. Zhang, D. Wu, and M. Guizani, "Deep features learning for medical image analysis with convolutional autoencoder neural network," *IEEE Trans. Big Data*, early access, Jun. 20, 2017, doi: 10.1109/TBDATA.2017.2717439.
[20] W. Pan, Y. Dong, and D. Wu, *Classification of Malaria-Infected Cells Using Deep Convolutional Neural Networks*. London, U.K.: IntechOpen, 2018, pp. 159–172.
[21] S. Rajaraman, S. Jaeger, and S. K. Antani, "Performance evaluation of deep neural ensembles toward malaria parasite detection in thin-blood smear images," *PeerJ*, vol. 7, p. e6977, May 2019.
[22] J. Lemley, S. Bazrafkan, and P. Corcoran, "Smart augmentation learning an optimal data augmentation strategy," *IEEE Access*, vol. 5, pp. 5858–5869, 2017.
[23] O. Bailo, D. Ham, and Y. M. Shin, "Red blood cell image generation for data augmentation using conditional generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Long Beach, CA, USA, Jun. 2019, pp. 1039–1048.
[24] W. Johnson and J. Lindenstrauss, "Extensions of Lipschitz mappings into a Hilbert space," *Contemp. Math.*, vol. 26, pp. 189–206, Jan. 1984.

[25] V. Athitsos, M. Potamias, P. Papapetrou, and G. Kollios, "Nearest neighbor retrieval using distance-based hashing," in *Proc. IEEE 24th Int. Conf. Data Eng.*, Apr. 2008, pp. 327–336.
[26] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," 2016, *arXiv:1612.08242*. [Online]. Available: http://arxiv.org/abs/1612.08242
[27] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: http://arxiv.org/abs/1804.02767
[28] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," 2017, *arXiv:1711.05225*. [Online]. Available: http://arxiv.org/abs/1711.05225

RABIAH AL-QUDAH received the B.Sc. degree in computer science from the Jordan University of Science and Technology and the M.Sc. degree in computer science from Concordia University, in 2018, where she is currently pursuing the Ph.D. degree in computer science.

She has seven years of experience in software and database development. She has published several articles in international conferences and journals. Her research interests include deep learning and data science. She was awarded the Best Paper Award in ICVISP 2019.

CHING Y. SUEN (Life Fellow, IEEE) received the M.S. degree in electronics from the University of Hong Kong, Hong Kong, in 1968, and the Ph.D. degree in man–computer communications from the University of British Columbia, Vancouver, BC, Canada, in 1972. In 1972, he joined the Department of Computer Science, Concordia University, Montreal, QC, Canada, where he was a Professor in 1979, the Chairman from 1980 to 1984, and the Associate Dean (Research) with the Faculty of Engineering and Computer Science. He is currently the Director of the Center for Pattern Recognition and Machine Intelligence, Concordia University. He has supervised 120 doctoral and master's students to completion and guided/hosted 100 long-term visiting scientists and professors. He has always been fascinated by letters and characters, ever since he started his Ph.D. research on teaching the computer to read multifont documents with a voice output for the blind. He has authored or coauthored six conference proceedings, 15 books, and more than 550 articles. He became a Fellow of the IAPR, in 1994, and the Academy of Sciences of the Royal Society of Canada, in 1995. He has served at numerous national and international professional societies as the President, the Vice President, the Governor, and the Director. He is the Founder of three conferences, such as the International Conference on Document Analysis and Recognition (ICDAR), the International Workshop/Conference on Frontiers in Handwriting Recognition, and Vision Interface, and has also organized numerous international conferences, including the International Conference on Pattern Recognition, ICDAR, ICFHR, and the International Conference on the Computer Processing of Oriental Languages. He was a recipient of numerous awards, including the Gold Medal from the University of Bari, Italy, in 2012, the IAPR ICDAR Award in 2005, the ITAC/NSERC National Award, in 1992, and the Concordia Fellow Award and the Concordia Lifetime Research Achievement Award, in 1998 and 2008. In 1997, he created the IAPR ICDAR Awards, to honor both young and established outstanding researchers in the field of document analysis and recognition. He was the Editor-in-Chief of *Pattern Recognition* for ten years and became the Emeritus EIC, in 2018.