**Artificial Intelligence in Medicine**

# DETECTION OF ABNORMALITIES IN BLOOD CELLS USING A REGION-BASED SEGMENTATION APPROACH AND SUPERVISED MACHINE LEARNING ALGORITHM

**NAGUEU DJAMBONG Lionel Perin***

Université de Yaoundé I

*perin.nagueu@facsciences-uy1.cm*

**WAKU KOUOMOU Jules**†

Université de Yaoundé I

*jules.waku@yahoo.fr*

**Hippolyte KENFACK TAPAMO***

Université de Yaoundé I

*htapamo@gmail.com*

**Jimbo H. Claver***

American International University Kuwait City, KUWAIT

*jimbo.maths@gmail.com*

**Abstract**

Screening (slide reading stage) is a manual human activity in cytology which consists of the inspection or analysis by the cytotechnician of all the cells present on a slide. Segmentation of blood cells is an important research question in hematology and other related fields. Since this activity is human-based, detection of abnormal cells becomes difficult. Nowadays, medical image processing has recently become a very important discipline for computer-aided diagnosis, in which many methods are applied to solve real problems. Our research work is in the field of computer-assisted diagnosis on blood images for the detection of abnormal cells. To this end, we propose a hybrid segmentation method to extract the correct shape from the nuclei to extract features and classify them using SVM and KNN binary classifiers. In order to evaluate the performance of hybrid segmentation and the choice of the classification model, we carried out a comparative study between our hybrid segmentation method followed by our SVM classification model and a segmentation method based on global thresholding followed by a KNN classification model. After this study, it appears from the experiments carried out on the 62 images of blood smears, that the SVM binary classification model gives us an accuracy of 97% for the hybrid segmentation and 57% in the global thresholding and 95 % for the KNN Classification Model. As our dataset was not balanced, we evaluated precision, recall, F1 score and cross validation with the Stratified K-Fold cross validation algorithm of each of these segmentation methods and classification models. We obtain respectively: 93.75%; 98.712% and 99% for hybrid segmentation reflecting its effectiveness compared to global fixed threshold segmentation and KNN classification model. To evaluate the performance of these models we obtained the following results: 77% of mean accuracy in the SVM and 61% of mean accuracy in the KNN, 84% of mean test accuracy in the SVM and 74% mean test accuracy in the KNN making the best performing SVM model

**Keywords:** Cytology, image segmentation, classification, model, computer-assisted diagnosis, cross validation.

# Introduction

In pathological anatomy there are two types of examination: histological and cytological. Both can be used to make a diagnosis, but sometimes only one of them can be decisive. Histology is the observation of the microscopic section of a tissue, whereas cytology examines a spread of isolated cells from samples. In our context we will only be interested in cytological analysis, which is the detection of abnormal or suspicious cells in order to establish a reliable and valid diagnosis [1]. The analysis of blood cells in microscopic images can provide useful information about the health of patients. This reading step, called screening, is a manual activity that consists of a visual inspection and analysis by the cyto-technician of all the cells present on a slide, with the aim of detecting abnormal or suspicious cells in order to establish a diagnosis [1]. After several investigations carried out in health centres, clinics and hospitals in Yaoundé, we noticed that this analysis is carried out manually using ordinary and not digital microscopes; this is due to the lack of high quality materials. This analysis is of capital interest because the diagnosis depends on the good recognition of abnormal or suspect cells. However, this is difficult and always a very time-consuming process that requires a lot of concentration, time and skill, sometimes leading to errors when these results are left to the cyto technician. Since this heavy burden rests solely on the pathologist, we need to solve the problem of automatic detection of abnormal or suspicious cells in these blood smear samples. Thus our research question will be how to classify cell images using segmentation methods to highlight the shape of the nuclei. We will first recall the different works that have been carried out on this problem, then we will present our methodology to solve the problem and finally we will present our results from this methodology.

# 1 State of Art

Segmentation is a vast subject of study and is one of the major themes of digital imaging. In this respect, many publications report on segmentations [2] [3]. The choice of an appropriate type of segmentation remains an open debate. Indeed, to correctly validate a segmentation of natural objects, as in medical imaging, one needs to have the ground truth. This is not obvious in the case of segmentation, as it is difficult to define precisely where objects start and stop in an image. Therefore, there is not only one way, but several possibilities to segment an image, and they are very often subjective. Similarly, depending on what we want to segment, certain techniques will be more likely to achieve this. Thus the various works that have been carried out for the segmentation and classification of blood cells will be limited to the methods of the region-based approach.

**Khin Yadanar Win et al** [4] proposed an Otsu thresholding method to automatically segment cell nuclei in pleural fluid cytology images. In the proposed method, the original image is first enhanced using a median filter, then the enhanced image is converted to the l*a*b*(where l* is the lightness derived from the luminance of the surface and the two parameters a* and b* express the deviation of the color from that of a gray surface of the same lightness) colour space and the l* and b* components are extracted. Cell nuclei are segmented using Otsu thresholding as a binary image. Subsequently, morphological operations are used to remove unwanted features and reconstruct the segmented image in colour.

**Olivier Lezoray et al** [1] developed a semi-automatic system for detecting screening errors based on cytoplasmic and nuclear strategies. They use the marker constrained watershed method (colours). Since their strategies are performed on colour images and there is a plethora of colour spaces, they limited themselves to the RGB, HSL and $l^*$, $u^*$ and $v^*$. The method developed is carried out as follows: Extraction of markers from the regions to be segmented (cytoplasmic and nuclear), determination of the image on which the L.P.E. is calculated and calculation of the L.P.E. (using the morphological gradient as a potential function).

**Minal D.Joshi et al** [5] proposed a system for the detection of acute leukaemia using algorithms for the segmentation and classification of blood cells (leukocytes). The proposed system is as follows:

- **Segmentation :** performed here using an automatic Otsu thresholding method to segment blood cell nuclei and extract their characteristics.
- **Extraction of characteristics:** consisted here in transforming the input data into a different set of features. Thus three features were extracted from the binary image obtained from the segmentation; the surface, the perimeter and the circularity or compactness.
- **Classification:** based on the features extracted from the segmentation, the classifier classifies the lymphocyte cells as blast or normal. **Mina D.Joshi et al** used KNN (K-nearest neighbour) classification which is a non-paerometric classification method. This method is used to classify blast cells from normal white blood cells.

**Muhammad Sajjad et al** [6] proposed a system for segmentation and classification of leukocytes in microscopic blood smears. In their system the nucleus is first segmented, then the features

of textures, geometries and statistics are extracted and finally the features are passed to a multi-class classification to detect the different sub-classes of leukocytes. Thus the system is subdivided into three steps: segmentation, feature extraction and classification.

— **Segmentation :** The authors use a K-means segmentation algorithm which is an algorithm where the speed depends on the number of K clusters. To segment white blood cells from blood smears, they used k=4(clusters) in which the original image was first enhanced and converted to HSL colour space and then extracted the leukocytes.

— **Extraction of characteristics:** After segmentation, they extracted the characteristics (geometric, statistical, textural) and labels of the segmented kernels and classified them.

— **Classification:** The authors used a set of multi-class SVMs (EMC-SVM) for the classification of leukocytes into five classes. This is due to the diversity of blood smear images for which training a single classifier is impractical due to its limited performance. Their proposed SMC-SVM was designed to classify white blood cells into different classes namely: Lymphocytes, Basophils, Neutrophils, Eosinophils and Monocytes.

# 2 Methodology

In this section we propose an automatic model for detecting abnormal cells in blood smear images. This model is subdivided into three different steps, namely: **the segmentation** used to locate nuclei in our blood images; **extraction of characteristics** segmented cores and the **classification** of these nuclei to determine abnormal or normal cells.

## 2.1 Segmentation

Segmentation in our work will allow us to highlight segments that correspond to objects, parts of objects or groups of objects that appear in an image. In our case, the objects to be highlighted are cell nuclei from blood tests. To cope with the segmentation of nuclei, a hybrid segmentation process is proposed. It is based on two segmentation techniques: global thresholding and morphological opening based on dilation and erosion-morphological operations. Our segmentation technique is given by the following formula (1) [7].

$$Segmentation = \begin{cases} (1)\text{Fixed overall threshold} \\ (2)\text{Morphological openness} \end{cases} \quad (1)$$

### 2.1.1 Segmentation by fixed-threshold global thresholding

The segmentation procedure starts with a global thresholding technique that is applied to the colour image $I_{RVB}$ which produces a binary image $I_{BW}$ in which the nuclei are visible. This thresholding consists in comparing the grey level of each pixel $x_i$ of the image with a fixed global threshold T. The algorithm 1 below represents the Fixed overall threshold

**Algorithm 1 Fixed overall threshold**

Enter: I= Colour image of blood smear
Out: $I_{seg}$ = Segmented image
For i ranging from 1 to height(I)
For j ranging from 1 to width(I)
if $x_i(i,j) \geq T$ then
$I_{BW}(i,j) = 1$
if $x_i(i,j) < T$
$I_{BW}(i,j) = 0$
endif
endfor
endfor
$I_{seg} = I_{BW}$

Where T was chosen manually after several tests carried out on the images in order to have a clearer cell nucleus shape, $I_{seg}$ is the segmented image and $x_i$ the value of each pixel. The threshold T is the integer value between 0 and 255 and $I_{BW}$ represents the new pixel value. After a thresholding operation, further processing is required to remove noise from the previously segmented image. This task is performed using a morphological operation called morphological opening. In addition, the segmented kernels in the image are black on a white background and it is more intuitive to work with white objects on a black background. if This task is performed by the image complement. Then the holes in the segmented areas are filled.

### 2.1.2 Morphological Segmentation

Mathematical morphology is a mathematical and computational technique of analysis that is related to algebra, topology and probability. Its basic principle is to compare an unknown shape with a known reference shape, called a structuring element. This element scans the whole set and allows at each point to make a comparison through relations such as union, intersection, inclusion and complementation. The mathematical morphology approach aims to determine the characteristics of an object, simplify the image by removing certain geometric structures, separate glued objects and

compare two shapes using the structuring element. This theory uses two basic operations which are erosion and dilation. To implement our segmentation by morphological opening it is necessary to have a base on some notion of mathematical morphology and among these notions we have:

— **Structuring element:** Which is a mask of any shape whose elements form a pattern. Let B be a subset of E, called a structuring element. If x is an element of E, then we define a set $B_x$, the displacement of B at each point x in space E:

$$B_x = \{b + x |\ b\ \in\ B\}$$

We introduce the symmetry of B noted $B_s$: $B_s = \{-b, \forall\ b\ \in\ B\}$
If the structuring element is symmetrical, we have : $B_s = B$

— **Morphological dilatation:** A morphological dilation consists in moving the structuring element on each pixel of the image, and looking if the structuring element touches the structure of interest. Let X be a subset of E, the morphological dilation of X by a structuring element B, note $\delta B(X)$ is defined as the Minkowski sum:
$\delta B(X) = X \oplus B = \{x + b |\ b\ \epsilon\ B\}, x\ \in\ \mathbb{R}^2 = \bigcup_{x \in \mathbb{R}^2}$
algorithm 2 is a Morphological dilatation.

**Algorithm 2 dilation of a binary image by a structuring element**

Enter: I= Segmented image
Out: $I_{dilate}$= Expanded images
For i ranging from 1 to height(I)
For j ranging from 1 to width(I)
if $I(i, j) = 0$ and different from one of the 8 neighbours then
$I_{dilate}(i, j) = 1$
endif
endfor
endfor $retourner I_{dilate}$

— **Morphological erosion**: It consists of searching for all pixels for which the structuring element centred on that pixel touches the outside of the structure. The morphological erosion of a set X by a structuring element B, noted $\epsilon B(X)$ is the set of points x in space for which $B_x$ is contained in X:
$\epsilon B(X) = X \ominus B = \{x \in \mathbb{R}^2 \subset X\}$.
algorithm 3 is Morphological erosion.

**Algorithm 3 Erosion of a binary image by a structuring element**

Enter: I= Segmented imaged
Out: $I_{erode}$= Eroded image
For i ranging from 1 to height(I)
For j ranging from 1 to width(I)
if $I(i, j) = 1$ and different from one of the 8 neighbours then
$I_{erode}(i, j) = 0$
endif
endfor
endfor
$retourner I_{erode}$

— **Morphological openness**: The morphological opening of a set X noted $X \circ B$, is erosion $B_s$ followed by a dilation with B:
$X \circ B = \delta B_s((\epsilon B(X)))$
With the use of the symmetrical element, this amounts to performing both operations with the same kernel. The opening is therefore defined as:
$X \circ B = \delta B((\epsilon B(X)))$

Morphological opening is applied to the binary image obtained in the previous segmentation step, in order to remove the noise and to break the union points that constitute the overlaps between the cells. The structuring element used for this operation is a disc of radius R = 2, this shape as well as the value of the radius is justified respectively by the morphology and the size of the nuclei that we want to preserve with respect to the noise. The disc allows the contour of the nuclei to be smoothed.

The segmentation method we used is shown in Figure 1.

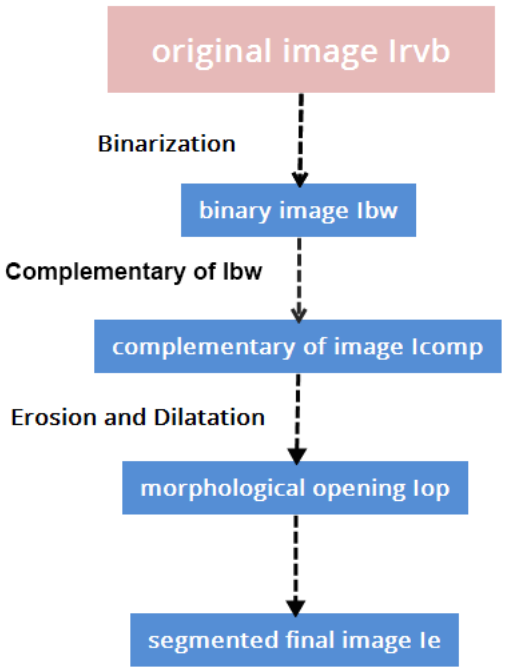



Figure 1 – Segmentation method

The aim of our segmentation will be to identify the fragments of the cell image containing the nuclei.

## 2.2 Extraction of characteristics

The quantification of the extracted features will help us to differentiate between abnormal and normal cells. For each isolated nucleus, the selected features can be divided into three groups: morphological features related to the size and shape of the nuclei, intensity features that provide information about the intensity histogram of the pixels located in the nuclei, and texture features giving information related to the intensity variation of a surface. In the case of our method we will only use morphological and textural characteristics.

### 2.2.1 Morphological characteristics

The morphological features of the nuclei that were extracted from our segmented images are as follows:

- **Area(A):** This is the number of pixels covering the surface of each detected core

$$A = \sum_{i=1}^{n} \sum_{j=1}^{m} B(i,j)$$

  B is the segmented image of dimension i×j.
- **Perimeter:** This is the number of pixels outlining each core.
- **Circularity or compactness:** It is a dimensionless parameter that changes with surface irregularities, and is calculated from the perimeter (P) and the area (S). The circularity or compactness (C) is calculated as follows:

$$C = \frac{4\pi S}{P^2}$$

### 2.2.2 Textural characteristics

Texture is a connected set of pixels that occurs repeatedly in an image. Texture analysis techniques are based on the Grey Level Co-occurrence Matrix (GLCM). The co-occurrence matrix P is of dimension N×N. In other words, each element of P indicates the number of times a pixel with grey level value i arrives shifted by a given distance to a pixel with value j. Here we average four GLCM features determined for shifts corresponding to $0°, 45°$ et $135°$ using eight levels of grey. After segmenting our cell nuclei the textural features we extracted are presented below:

- **Contrast :** Measures the intensity of contrast between a pixel and its neighbours.

$$Contrast = \sum_{(i,j)=1}^{N} |i-j| p(i,j)$$

- **Correlation:** Measurement of the grey level dependency of the image.

$$Correlation = \sum_{(i,j)=1}^{N} \frac{(i-\mu_i)(j-\mu_j)p(i,j)}{\sigma_i \sigma_j}$$

- **Energy or uniformity:** Is the sum of the squares of the elements in the grey level co-occurrence matrix (GLCM).

$$Energy = \sum_{(i,j)=1}^{N} p(i,j)^2$$

- **Homogeneity:** Measures the proximity of the distribution of the GLCM elements to the diagonal of the GLCM.

$$Homogeneity = \sum_{(i,j)=1}^{N} \frac{p(i,j)}{1+|i-j|}$$

# 3 SVM classification

Classification is considered to be the last step in a computer-aided diagnostic system. It uses the result of feature extraction to decide on the pathological nature of the images. The concept of classification means assigning a label to samples in a database using a number of features. To classify our cells into abnormal and normal we will use a classification by SVM (support vector machine) which is a supervised classification algorithm that consists

of determining a separator between two classes of point sets, whose margin on either side of this separator is maximum. The margin is defined with respect to the support vectors, which are the points closest to the hyperplane. to the hyperplane. Since our data is non-linear, we used a polynomial kernel to make it linear and applied our binary classification [8][9]. this core is defined by the following formula (2)

$$K(x, y) = (1 + x^T y)^d \qquad (2)$$

if d=1 we speak of a linear core.

# 4 KNN classification

In machine learning, the K-nearest neighbour (KNN) method is a supervised learning method. In this method, we have a training database consisting of N "input-output" pairs. To estimate the output associated with a new input x, the K-nearest neighbour method consists of taking into account the k training samples whose input is closer to the new input x, according to a well-defined distance. Since this algorithm is based on distance, a normalisation can improve its accuracy.

This method has two forms: classification and pattern recognition. In our context we will only focus on classification. In a classification problem, we will obtain the most represented class among the k neighbours associated with the k closest inputs to the new input [10]. The KNN algorithm will proceed as follows:

- **Goals**: assign a class to a new instance
- **Data**: a sample of m records classified (x, c(x))
- **Enter**: a y record
  - Determine the k nearest records of y
  - combine the classes of these k examples into a class c
- The class of y is c(y)=c

# 5 Cross validation

Cross-Validation is a method for testing the performance of a machine learning predictive model. However, it is important to ensure the accuracy of the model's predictions in production. To do this, it is necessary to validate the model. The validation process involves deciding whether numerical results quantifying hypothetical relationships between variables are acceptable as descriptions of the data. So to evaluate our model with cross-validation several algorithms are proposed and among these algorithms we have:

- **K-Fold Cross Validation:** As there is never enough data to train the model, removing a part of it for validation poses a problem of underfitting. By reducing the training data, we risk losing important patterns/ trends in data set, which in turn increases error induced by bias. So, what we require is a method that provides sample data for training the model and also leaves ample data for validation. K Fold cross validation does exactly that. In K Fold cross validation, the data is divided into k subsets. Now the holdout method is repeated k times, such that each time, one of the k subsets is used as the test set/ validation set and the other k-1 subsets are put together to form a training set. The error estimation is averaged over all k trials to get total effectiveness of our model. As can be seen, every data point gets to be in a validation set exactly once, and gets to be in a training set k-1 times. This significantly reduces bias as we are using most of the data for fitting, and also significantly reduces variance as most of the data is also being used in validation set. Interchanging the training and test sets also adds to the effectiveness of this method. As a general rule and empirical evidence, K = 5 or 10 is generally preferred, but nothing's fixed and it can take any value.
- **Stratified K-Fold Cross Validation:** In some cases, there may be a large imbalance in the response variables. For example, in dataset concerning price of houses, there might be large number of houses having high price, or in case of classification, there might be several times more negative samples than positive samples. For such problems, a slight variation in the K Fold cross validation technique is made, such that each fold contains approximately the same percentage of samples of each target class as the complete set, or in case of prediction problems, the mean response value is approximately equal in all the folds. This variation is also known as Stratified K Fold.
- **Leave-P-Out Cross Validation:** This approach leaves p data points out of training data, i.e. if there are n data points in the original sample then, n-p samples are used to train the model and p points are used as the validation set. This is repeated for all combinations in which original sample can be separated this way, and then the error is averaged for all trials, to give overall effectiveness. This method is exhaustive in the sense that it needs to train and validate the model for all possible combinations, and for moderately large p, it can become computationally infeasible. A particular case of this method is when p = 1. This is known as

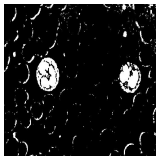
Leave one out cross validation. This method is generally preferred over the previous one because it does not suffer from the intensive computation, as number of possible combinations is equal to number of data points in original sample or n.

# 6 Experiments and results

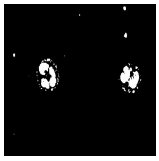
After defining our methodology as presented above, we will present the results obtained from the image segmentation model and the classification of blood smears. For the classification of abnormal and normal cells we used two supervised classification algorithms (binary SVM and KNN) and compared their respective performances.

## 6.1 Segmentation

To segment our cell images well we need to have a good threshold value and a good structuring element. Thus, to choose the right value of the threshold (T) and of our structuring element (R), we conducted experiments with several values of T and E. In view of our observations, we realised that the value T=29 and R=2 (disc of radius 2) allows us to obtain the exact shape of the cell nuclei. The figure 2a 2b 2c The following presents the different experiments on the choice of the threshold and the structuring element.



(a) overall fixed threshold T=12



(b) overall fixed threshold T=29

(c) Structural element R=2

The table below shows the results of the experiments with our segmentation method.

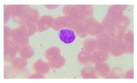
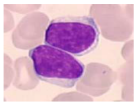
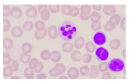
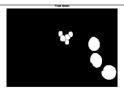
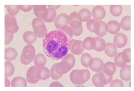
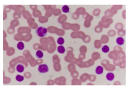
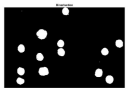
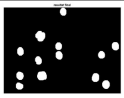
| Original image | Binarisation | thresholding and morphological opening |
|---|---|---|
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

Table 1 – Result of our segmentation method

We can see from the table 1 that:

- For the original image containing a single cell nucleus (first row and first column), the global thresholding segmentation manages to bring out the shape of the nucleus perfectly.
- For the original images with several cell nuclei, the global thresholding with fixed threshold does not manage to bring out exactly the shape of the different nuclei present in the images (original image of the second line). For this reason, we use a morphological aperture that allows us to clearly show the shape of the nuclei.

## 6.2 Extraction of morphological and textural characteristics

The extraction of features from the segmented blood smear images is an important phase, as our classification will depend on these features. To perform cell classification, we extracted morphological (perimeter, area and circularity) and textural (contrast, correlation, energy and homogeneity) features from the segmented nuclei. This extraction was done on two segmentations (global thresholding and our hybrid segmentation).

## 6.3 Classification

### 6.3.1 Evaluation metrics

To evaluate the performance of our SVM classifier, we used several metrics, namely.

- **Precision:** It represents the proportion of correctly predicted normal images. It is defined by the following formula:

$$\text{accuracy} = \frac{VP}{VP+FP}$$

- **Recall**: For a given class, it represents the proportion of images correctly predicted. It is defined by the following formula:

$$Recall = \frac{VP}{VP + FN}$$

- **F1-score**: It is the harmonic mean between precision and recall. It gives the precision of the classifier and the most suitable for unbalanced data sets. It is defined by the following formula:

$$F1 - score = 2 \times \frac{(precision \times recall)}{(precision + recall)}$$

- **Accuracy:** which is defined here as the proportion of correct predictions. Its formula is shown below:

$$Accuracy = \frac{VP + VN}{VP + VN + FP + FN}$$

With VP, FP, VN and FN representing respectively:

⋆ True positive: a normal nucleus classified as normal.
⋆ True negative: an abnormal nucleus classified as abnormal.
⋆ False positive: a normal nucleus classified as abnormal.
⋆ False negative: an abnormal nucleus classified as normal.

#### 6.3.2 SVM classification results

The model was trained with 62 images of segmented blood smears consisting of several varieties of nuclei, namely: lymphocytes, basophils, neutrophils, eosinophils and monocytes. The aim of this training is to determine whether we observe abnormalities in these nuclei or not. Thus, we have separated our data set into 67% for training and 46% for the tests which gives us a total of 113%.
To train our model we had a dataset of 62 images distributed as follows: the table2 presents the training results of the binary SVM classifier based on the features extracted after segmentation of the global thresholding.

| Classes | Precision | Recall | F1-score |
|---|---|---|---|
| Normal | 75% | 17% | 27% |
| Abnormal | 56% | 95% | 70% |

Table 2 – Classification result of the binary SVM classifier after global thresholding

The dataset used for the classification after global thresholding is the same for the classification after hybrid segmentation, the table 3 presents the training results of the binary SVM classifier based on the features extracted after the hybrid segmentation.

| Classes | Precision | Recall | F1-score |
|---|---|---|---|
| Normal | 88% | 100% | 93% |
| Abnormal | 100% | 97% | 99% |

Table 3 – Classification result by the SVM binary classifier after hybrid segmentation

The table 4 presents a comparison of the results obtained from training the binary SVM classifier after the segmentation of the global fixed threshold and hybrid thresholding.

| Metrics | threshold | **Hybrid** |
|---|---|---|
| Precision | 65, 44% | **93,75**% |
| Recall | 55, 83% | **98,71**% |
| F1-score | 70% | **99**% |
| Accuracy | 57, 84% | **97,82**% |

Table 4 – Comparison of the two methods' classification

#### 6.3.3 Interpretation of SVM classification results

The table 4 presents a comparative study of the classification of the two segmentation methods. To evaluate the performance of the chosen hybrid segmentation, we performed a comparative study with another segmentation method based on global thresholding. After this study, the experiments carried out on the 62 blood smear images show that the binary SVM classification model gives us an accuracy of 97% for the hybrid segmentation 57% in the global thresholding. As our dataset is not balanced, we have evaluated **the precision**, **the recall** et **the F1-score** on each of these methods. We obtain respectively: 93,75%; 98,71% et 99% for hybrid segmentation that reflect its effectiveness compared to global fixed threshold segmentation.

#### 6.3.4 KNN classification results

The model was trained with 62 images of segmented blood smears consisting of several varieties of nuclei, namely: lymphocytes, basophils, neutrophils, eosinophils and monocytes. The aim of this training is to determine whether we observe abnormalities in these nuclei or not. Thus, we have separated our data set into 67% for training and 46% for the tests and validations.
To train our model we had a dataset of 62 images distributed as follows: the table2 presents the training results of the KNN classifier based on the

features extracted after segmentation of the global thresholding.

| Classes | Precision | Recall | F1-score |
|---|---|---|---|
| Normal | 71% | 76% | 73% |
| Abnormal | 53% | 47% | 50% |

Table 5 – Classification result of the KNN classifier after global thresholding.

The dataset used for the classification after global thresholding is the same for the classification after hybrid segmentation, the table 6presents the training results of KNN classifier based on the features extracted after the hybrid segmentation.

| Classes | Precision | Recall | F1-score |
|---|---|---|---|
| Normal | 100% | 95% | 98% |
| Abnormal | 67% | 100% | 80% |

Table 6 – Result of the classification by the KNN classifier after the hybrid segmentation.

#### 6.3.5 Interpretation of the KNN classification results

The table 4 presents a comparison of the results obtained from training the KNN classifier after the segmentation of the global fixed threshold and hybrid thresholding.

| Metrics | threshold | **Hybrid** |
|---|---|---|
| Precision | 62, 15% | **83,33**% |
| Recall | 61, 46% | **97,61**% |
| F1-score | 73% | **98**% |
| Accuracy | 65, 21% | **95,65**% |

Table 7 – Comparison of the classification of the two methods in the KNN classification.

The table 7 presents a comparative study of the KNN classification of the two segmentation methods. To evaluate the performance of the chosen hybrid segmentation, we performed a comparative study with another segmentation method based on global thresholding. After this study, it appears from the experiments done on the 62 blood smear images that the KNN classification model gives us an accuracy of 95,65% for hybrid segmentation and 65% dans le seuillage global à seuil fixe. As our dataset was not balanced we evaluated **the precision**, **the recall** et **the F1-score** on each of these methods. We obtain respectively: 83,33%; 97,61% et 98% for hybrid segmentation that reflect its effectiveness compared to global fixed threshold segmentation.

#### 6.3.6 Comparison between the SVM and KNN classification models

After different studies done on SVM and KNN classification, we came out with a comparative table of these two classification models defined in the table. 8

| Metrics | **SVM** | KNN |
|---|---|---|
| Precision | **93,75**% | 83,33% |
| Recall | **98,71**% | 97,61% |
| F1-score | **99**% | 98% |
| Accuracy | **97,21**% | 95,65% |

Table 8 – Comparison of the two classification models SVM and KNN.

The table 8 presents a comparative study of two classification models, namely; SVM and KNN. In order to choose the best classification model for our cellular images, we have based ourselves on the different metrics studied above (precision, recall, accuracy and F1-score).After studying these different metrics, it appears from this table that the SVM classification model gives us an accuracy of **97,21**% and the KNN classification model gives us an accuracy of 95,65%. Given that our data is unbalanced, we evaluated the performance of our classification models on other metrics such as: precision, recall and F1-score, which gave us respective results; **93,75**%, **98,71**%, **99**% for SVM and 83,33%, 97,61%, 98% for KNN. From this study we can say that the SVM classification model provides better performance compared to the KNN classification model, hence our SVM classification model is the most suitable to classify our different blood cells into abnormal and normal.

#### 6.3.7 Evaluation of the proposed models

In order to measure the performance of our different proposed models (SVM and KNN) we opted for cross-validation with the Stratified K-Fold Cross Validation algorithm because our dataset was not balanced (37 for abnormal cells and 25 for normal cells). The purpose of this algorithm was initially to balance our dataset in order to better evaluate the performance of our models. For the performance evaluation we used the same dataset described in the previous section, and after performing a Stratified K-Fold Cross Validation with 3 split (k=3 ) on our respective models (SVM and KNN) we have obtained the results present under the table **??** below:

| Stratified K-Fold | **SVM** | KNN |
|---|---|---|
| First iteration | 61,9% | **66,7**% |
| Second iteration | **85**% | 60% |
| Third iteration | **85**% | 55% |
| Mean accuracy | **77**% | 61% |
| Mean test accuracy | **84**% | 74% |

Table 9 – Performance of two classification models SVM and KNN.

This table presents a cross-validation using the Stratified K-Fold algorithm cross validation on our two classification models, namely; SVM and KNN. In order to choose the classification model with the best performance, we based ourselves on the different iterations, mean accuracy and mean test accuracy. After validation of these models, it appears from this table that the SVM classification model gives us a mean accuracy of **77**% and a mean test accuracy of **84**%. As for the KNN classification model, it gives us mean accuracy of 61% and a mean test accuracy of **74**%. Which leads us to conclude that our SVM classification model has better performance than that of KNN and is the most suitable for classifying our different blood cells into abnormal and normal.

# 7 Conclusion

We proposed a hybrid segmentation method consisting of global thresholding with morphological opening and classification using binary SVM and KNN algorithm as classifier. This method consists of first extracting the shapes of the nuclei, then extracting their morphological characteristics and finally passing them through the binary classifier SVM and KNN. Then we evaluated the performances of the hybrid segmentation chosen, by carrying out a comparative study with another method of segmentation based on the global thresholding, then we also proceeded to an evaluation of the performances of our two classification models by using a cross validation. After this study, it appears from the experiments carried out on the 62 images of blood smears, that the binary model of SVM classification gives us an accuracy of 97% for the hybrid segmentation and 57% in the global thresholding, then 97% for the SVM classification model and 95% for the KNN classification model. As our dataset was not balanced, we assessed the precision, recall, and F1 score of each of these methods. We obtain respectively: 93.75%; 98.71% and 99% for hybrid segmentation reflecting its effectiveness compared to fixed threshold global segmentation and the choice of our SVM classification model. To evaluate the two models studied, we opted for a cross validation with the Stratified K-Fold cross validation algorithm since our dataset was not balanced, after this validation we obtained the following results: 77% of mean accuracy in the SVM and 61% of mean accuracy in the KNN, 84% of mean test accuracy in the SVM and 74% mean test accuracy in the KNN thus making the SVM model the most efficient. Manual feature extraction and non-linearity of the used dataset can reduce the performance of our method. In future work, we will explore convolutional neural networks for automatic feature extraction and classification of abnormal cells. In addition, we plan to apply our method to the detection of hematological cancers and tumors present on brain images.

**Declaration of Competing interest**

There is no conflict of interest in this work.

**Data Availability**

Data are available on request.

**Acknowledgement**

The authors would like to express their sincere thanks to the associate editor and anonymous reviewers for their constructive comments and suggestions to improve the quality of this manuscript.

# References

[1] Olivier Lezoray, Abderrahim Elmoataz: *Colour Image Segmentation; application to serous cytology for computerised cell sorting*, *HAL Id: hal-00960829*, March 2014.

[2] Abderrahmane Kefali, Toufik Sari et Mokhtar Sellami: *Evaluation of several image thresholding techniques for Ancient Arabic documents Colour Image Segmentation*; 2009

[3] Allem Alima: *Realization of a segmentation approach: Application on cytological images* , page:46-47.

[4] Khin Yadanar Win et Somsak Choomchuay: *Automated segmentation of cell nuclei in pleural fluid cytology images using OTSU thresholding*, *IEE*, 2017

[5] Minal D.Joshi: *Segmentation and classification of white blood cells for the detection of acute leukaemia*,2013

[6] Muhammad Sajjad et al: *Classification and segmentation of*

*leukocytes from microscopic blood smears; a resource-sensitive health service in smart cities*, 28 March 2017.

[7] Vincent Djoum:*Memorandum on Image Segmentation and Application to Breast Cancer Diagnosis*, 2021

[8] Ali Ghodsi: *Hard margin support vector machine*, October 2015

[9] Ali Ghodsi: *soft margin support vector machine*, October 2015.

[10] Chloé-Agathe Azencott: *Book: introduction to Machine Learning*, March 2017