Sickle-cell disease diagnosis support selecting the most appropriate machine learning method: Towards a general and interpretable approach for cell morphology analysis from microscopy images.

Nataša Petrović[c], Gabriel Moyà-Alcover[c], Antoni Jaume-i-Capó[c,*], Manuel González-Hidalgo[a,b]

[a]*SCOPIA research group. University of the Balearic Islands. Dpt. of Mathematics and Computer Science. Crta. Valldemossa, km 7.5, E-07122 Palma, Spain*
[b]*Health Research Institute of the Balearic Islands (IdISBa), E-07010 Palma, Spain*
[c]*UGiVIA research group. University of the Balearic Islands. Dpt. of Mathematics and Computer Science. Crta. Valldemossa, km 7.5, E-07122 Palma, Spain*

Abstract

In this work we propose an approach to select the classification method and features, based on the state-of-the-art, with best performance for diagnostic support through peripheral blood smear images of red blood cells. In our case we used samples of patients with sickle-cell disease which can be generalized for other study cases. To trust the behavior of the proposed system, we also analyzed the interpretability. We pre-processed and segmented microscopic images, to ensure high feature quality. We applied the methods used in the literature to extract the features from blood cells and the machine learning methods to classify their morphology. Next, we searched for their best parameters from the resulting data in the feature extraction phase. Then, we found the best parameters for every classifier using Randomized and Grid search. For the sake of scientific progress, we published parameters for each classifier, the implemented code library, the confusion matrices with the raw data, and we used the public erythrocytesIDB dataset for validation. We also defined how to select the most important features for classification to decrease the complexity and the training time, and for interpretability purpose in opaque models. Finally, comparing the best performing classification methods with the state-of-the-art, we obtained better results even with interpretable model classifiers.

Keywords: Red Blood Cell, Sickle-cell Disease, Microscopy Image, Machine Learning, Interpretability, Morphology Analysis

1. Introduction

In sickle-cell Disease (SCD), red blood cells (RBCs) have the shape of a sickle or half-moon instead of the smooth, circular shape as normal cells have. Because of the irregular shape, the cells can be stuck in the blood vessels, which causes blockages of the blood flow in the body. Lack of elasticity and their irregular shape lead to

---

*Corresponding author
   Email addresses: npe785@uib.es (Nataša Petrović), gabriel.moya@uib.es (Gabriel Moyà-Alcover), antoni.jaume@uib.es (Antoni Jaume-i-Capó), manuel.gonzalez@uib.es (Manuel González-Hidalgo)

shortened life of sickle erythrocytes and subsequent anaemia, often called sickle-cell anaemia. The WHO document "Sickle-cell disease and other haemoglobin disorders" [1] indicates that in 2011 around a 5% of the world's population carries trait genes for hemoglobin disorders, mainly sickle cell disease and thalassemia. SCD is caused by inherited mutated haemoglobin genes from both parents. The document also indicates that the percentage of people who carry these genes is as high as 25% in some regions and over 300,000 babies with severe haemoglobin disorders are born each year. SCD is spread among people whose ancestors are from sub-Saharan Africa, India, Saudi Arabia and Mediterranean countries. Around the world [2, 3], SCD caused 112,900 deaths in 1990, 176,200 deaths in 2013 and 55,3000 deaths in 2016.

SCD cannot be cured but can be treated to ease the pain and prevent problems caused by this disease. One of the methods for diagnosing SCD is by observation of a patient's peripheral blood samples under a microscope and counting sickle cells, which is a tedious and time-consuming task. Morphological analysis of peripheral blood smears is a crucial diagnostic aid because offers important information for diagnosis [4]. Therefore, monitoring of these patients is an urgent requirement.

Machine learning (ML) algorithms have been used for solving different problems for medical image analysis [5, 6]. ML techniques are also used for analysis of microscopic blood smear images. These techniques make easier and cheaper solving tasks of diagnosing the different blood diseases such as malaria and different types of anemia and leukemia [7, 8]. Therefore, SCD is a great candidate for using ML algorithms as diagnosis support, so the medical experts could have more time to solve more complicated tasks.

A review about blood cells classification, focused on malignant lymphoid cells and blast cells, was published in [9]. There are three main steps in image processing systems that are able to recognize different classes of RBC in peripheral blood smear images: image segmentation and cell detection, feature extraction and/or selection, and finally classification. The two first steps will be addressed in subsequent sections, in next paragraphs we survey the most popular RBCs classification methods for images of peripheral blood smear samples.

Automatic classification is the process that tries to assign a set of features or descriptors to a specific class among a set of known classes defined as target. This is a well-established problem and dealt with machine learning techniques [10]. In RBCs classification the issue is assigning a given cell image its corresponding group.

Since the early work of Bacus et al. [11], much efforts has been made to design automatic classification methods of RBCs based on images. Among the most used classifiers for RBCs classification from peripheral blood smear samples, we found artificial neural networks (ANN), multilayer perceptron (MLP), Naïve Bayes (NB), support vector machines (SVM), k-nearest neighbors algorithm (kNN), and decision trees (DT). The most popular are, by far, the SVM, kNN and ANN classifiers.

One of the first ML methods used for the purpose previously described, was the classification and regression tree (CART) in 1994. That year, Wheeless et al. [12] used CART to develop a cell classifier in three target classes (normal, sickle and other abnormal cells). A total of 42 features were extracted, and circularity was selected as the sole feature needed for segregating cells into the target classes. After segmentation, cells touching the borders or cluster cells were

rejected, and individual detected cells were classified. The agreement between the automated classification and the classification provided by an expert was: 89% for normal disocytes, 73% for abnormal cells and 92% for sickle cells. The CART classifier would appear later in 2012 in the work of Das et al. [13].

Maity et al. [8] presented an ensemble rule-based decision-making approach for morphological classification of RBC. The decision about the abnormality was taken using multiple rule based expert systems. The rules were generated using a C4.5 DT algorithm. The proposed ensemble approach classified the RBCs in eight types of abnormal erythrocytes using a set of 41 shape features. Acharya et al. [14] developed a decision system with a set of rules to identify eleven categories of RBCs based on 8 geometrical categories. The authors used the K-medoids algorithm to obtain white blood cells (WBC) and granulometry analysis to separate the RBCs from WRCs. Clusters were not considered.

Akrimi et al. [15] used SVM to classify RBCs as either normal or abnormal for anemia diagnosis. Authors used an automated feature selection algorithm to obtain a set of 271 features, but they do not publish the details. Moreover, the kernel used in the experiments is not clear and the work is impossible to reproduce. The SVM classifier was used also in [16, 17].

The classification of RBCs using ANN were addressed in [16, 18–20]. After segmentation, Veluchamy et al. [18] used 27 features with an ANN (4 geometrical, 16 statistical and 7 moment invariant) to classify RBCs as normal or abnormal. Tomari et al. [19] used geometrical properties of cell (such us perimeter, area and seven Hu moment features) in combination with an ANN to classify individuals RBCs as normal or abnormal. Full range images were considered but clusters were not analysed in these works. Elsalamony [20] proposed an algorithm to detect 3 classes (sickle cell, elliptocytosis and unknown shapes). After a classification of healthy/unhealthy cells, using a set of 6 geometric features and three ANNs (one for each target class), the algorithm would be used to classify unhealthy cells. The output targets were measured with regard to the solidity and circularity.

Several works compared different classifiers [13, 16, 17, 21]. Das et al. [13] proposed a method for 6-class classification considering a set of 25 features and five supervised classifiers (Multivariate logistic regression (MLR), Radial basis function neural network (RBF-NN), MLP, NB, and CART). The selected significant features were ranked based on information gain and F-value. In this study overlapping cells were not considered. The better performance was obtained by the MLR classifier. Chen et al. [21] used a set of 14 features and 3 classifiers (J48 tree, NB, and DTNB rule) to identify abnormal erythrocytes using Weka software. The proposed technique recognized normal and abnormal erythrocytes using eight directional information, irregularity and differential value from chain codes technique. The parameters for each classifier algorithm were not provided. Only abnormal erythrocytes were used as input to conduct the classification of sickle cell anemia. Lofti et al. [16] used SVM, kNN and ANN to classify RBCs for iron deficiency anemia. The work used 33 determinant features which were either boundary or region descriptors, including 10 coefficients of Fourier descriptors and 7 Hu moments to classify 3 types of abnormal RBCs using SVM and kNN classifiers for each one of the 3 classes. Final classification was done by using maximum voting theory and samples classified in more than one class were eliminated. Clusters were not taking into account.

Rodrigues et al. [17] compared Naïve Bayes, kNN and SVM to classify a subset of isolated cells from the erythrocytesIDB dataset [1]. Clusters were not studied. 9 simple geometric shape features were considered and the best performance was obtained by the SVM classifier and the set of features excluding the perimeter.

Gual-Arnau et al. [22], as in [16, 17], proposed a method to automatically classify isolated erythrocytes cells as normal, sickle cells and cells with different deformation using the kNN classifier. The method used the generalized support function and some other functions related to or derived from it, to obtain a set of features related to the cells contour. Different sets of features were compared (based on the UNL-Fourier function, Fourier descriptors, Crofton's descriptor and the CSF-ESF shape descriptors). Sharma et al. [23] also proposed kNN classifier to detect three different types of distorted RBC (sickle cells, dacrocytes and elliptocytes) responsible of SCD and thalassemia, with an accuracy of 80.6% and sensitivity of 87.6%. Overlapping cells were not considered.

Besides the studies related to SCD from peripheral blood smear images, there is a large number of works related to the morphological analysis of blood cells from blood microscopic images. In [9], we can find a detailed account about ML methods for lymphoma and leukemia cells detection from peripheral blood smears. As was pointed by the authors, the most used methods in the classification of peripheral blood were ANN, SVM and the DT. Also in [24, 25] we can see how convolutional neural networks (CNN) were used to classify RBCs on raw microscopic images. A. Kihm et al. in [26] try to determine the shape of RBCs in images of microcapillary Poiseuille Flow using CNN's. In [27] using a deep convolutional neural network as feature extractor and later an error-correcting output codes classifier for the classification task, the authors classify RBCs in three classes with an accuracy of 92,06%. Later, the same authors in [28] analyze a set of deep learning models in combination with different scenarios to classify red blood cells. They conclude that their Model 2 Architecture is the most appropriate for the task at hand. In addition, they used the extracted features by the proposed deep models to train a multiclass SVM classifier, obtaining high accuracies in the data sets used. We should be note that the features extracted by the deep neural network models are very difficult to interpret and are not explainable. We can also consult a recent work on ML methods for the diagnosis of malaria by analyzing and interpreting microscopic images of thin blood smears [29].

Despite all the works consulted, most of the studies do not consider clusters of cells and cells in the border of the image, and as far as we are concerned, we have not found works that use random forest (RF), extra trees (ET) and gradient boosting (GB) for RBCs classification from peripheral blood smear images for SCD diagnosis. This is why the main objective of our work is to propose a methodology to select the classification method and features with best performance for diagnostic support through peripheral blood smear images of RBCs, in our case samples of patients with SCD which can be generalized for other study cases.

To accomplish this aim:

- We present all the features found in the literature that were used for cell classification tasks, specifically for morphology classification of blood cells.

---

[1] As we can see below, this dataset is available at `http://erythrocytesidb.uib.es/`

4

(a) normal   (b) sickle   (c) other

Figure 1: Examples of the three types of red blood cells to take into account in SCD diagnosis.

- We present classification methods reported in the literature as successful for blood cells morphology classification tasks.

- We determine the best parameters for all the selected classifiers to optimize the classification of sickle cells in blood smear images in a SCD case study.

- We determine the best performing classification methods based on the case study.

- We determine the most important features for the case study.

- We compare the best performing classification methods with other state-of-the-art algorithms.

To trust the behavior of diagnostic support systems, they must be able to explain their decisions and actions to human users [30]. Interpretable ML can cover the entire model, individual components as features, and the level of training [31]. As an important component of interpretable ML are the features, a key part of this work is to rank the features by their importance to guide our research towards interpretability of the proposed diagnostic support system.

This work is organized as follows: Section 2 presents the applied methods. Section 3 describes the conducted experiments. Section 4 presents the results and discusses these. The conclusion and future work are presented in Section 5.

2. Method

In this section, we propose a methodology to select the classification method and features with best performance for cell morphology analysis. Concretely, for SCD diagnostic support, there are three types of RBC to take in account based on their morphology. They can be normal (discocyte), deformed elongated (sickle cells), or cells with other deformations, as shown in Figure 1.

Figure 2 shows the methodology steps based on three main phases: image pre-processing and segmentation, feature extraction, and classification.
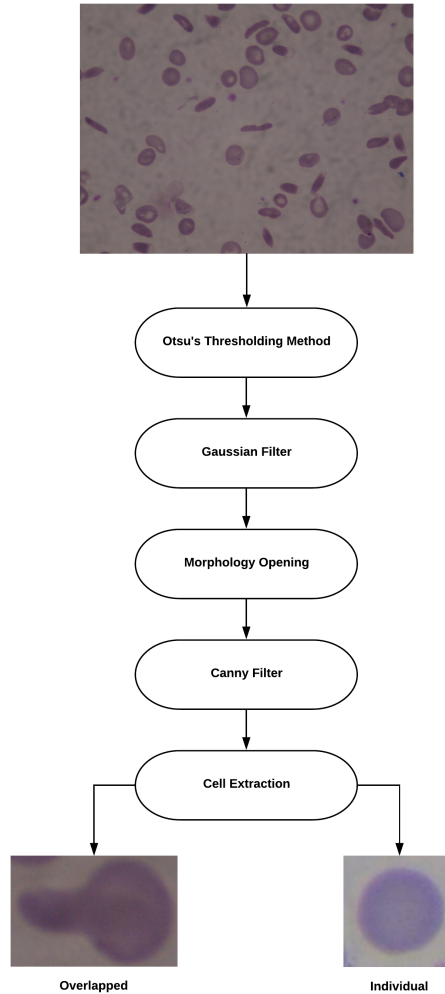
Figure 2: Methodology steps of a cell classification process: image preprocessing and segmentation (cell extraction), feature extraction and feature selection/reduction, and finally classification and comparison.

Figure 3: Image pre-processing and segmentation steps of the cell extraction process.

2.1. Image pre-processing

The first step to be taken before feature extraction is image pre-processing. This step is needed to reduce noise and ensure the quality of the extracted features [32]. Our main goal is to extract individual cells from microscopic images. An original image and the pre-processing steps taken are shown in Figure 3.

To prepare an image for cell extraction, we used standard methods for image processing. First, we binarized the original image using Otsu's thresholding. We chose this method because of the automatic optimal threshold value calculation it provides. Second, we reduced the image noise using a Gaussian filter and we eliminated small objects using morphology opening. This method performs erosion of an image followed by its dilation. All these steps lead to the most important one, which is edge detection. Edges of the image are recognized using the Canny filter, which uses a multi-stage algorithm to detect a wide range of edges in cell images [33]. As a result of this process, we can

extract the contours of every object in the image.

A problem we encountered was that some objects represent overlapping cells, as shown in Figure 3. In this case, an object's contour is shared by two or more cells. We solved this issue by applying a method based on ellipse adjustments and an algorithm for detecting notable points, as established by González-Hidalgo et al. [34].

As a result of this step we obtained a set of Regions of Interest (ROI) where each of them contains a cell.

2.2. Feature extraction

Feature extraction is the first step in any classification task using ML approaches, and is shown in Figure 2. Existing literature for blood cell detection in microscopic images has proposed many features [8, 35, 36]. We attempted to identify the three types of cells shown in Figure 1 in our classification task.

Regarding the identified cell types, we found the most-used features in the literature that can differentiate each type. Based on the state of the art, the extracted features belong to three categories: shape or geometry, color, and texture. We extracted forty-one shape features, eighteen color features, and sixty-two texture features from each cell for a total of 121 features to facilitate classification of the different cell types.

2.2.1. Shape features

Shape features provide measures of the cells based on their geometrical properties. We developed a shape feature extraction module that extracts forty-one different geometrical features from each segmented cell. The features and their descriptions are listed in Table 1. Let C be a closed curve enclosing a cell, and let O be the area enclosed by C. By $|C|$ we denote the number of pixels belonging to C, that is the length of the boundary cell, and by $|O|$ we denote the area enclosed by C (the area of the cell). Let $\mathbf{t}_i$ be the tangent line to each point $P_i \in C$ and d be the euclidean distance, then $d(\mathbf{t}_i, \mathbf{t}_j)$ measure the distance between two tangent lines to two points of C. At some point, we can also denote by ∂O the boundary of a region O.

Shape feature	Description	Equation	Ref
Perimeter	Number of boundary pixels	$|C|$	[9, 12–14, 16–18, 20, 24, 35, 37, 38]
Area	Number of pixels in the region inside the perimeter.	$|O|$	[8, 9, 11–14, 16–18, 20, 24, 35, 37, 38]
Min R (mR)	Maximum circle radius enclosed by the cell.		[8, 21, 38]
Max R (MR)	Minimum circle radius that encloses the cell.		[8, 21, 38]
Max feret diameter (MFD)	The maximum distance between parallel tangents to the boundary of the cell.	$\max_{P_i, P_j \in C} d(\mathbf{t}_i, \mathbf{t}_j)$	[8, 24, 38]
Min feret diameter (mFD)	The minimum distance between parallel tangents to the boundary of the cell.	$\min_{P_i, P_j \in C} d(\mathbf{t}_i, \mathbf{t}_j)$	[8, 24, 38]
Major axis (MA)	The length of the major axis of the ellipse that fits O by least square method.		[8, 12, 13, 16, 23, 24, 35, 37, 38]
Major axis angle (αMA)	The orientation of the major axis.		[8, 24, 35, 38]
Minor axis (mA)	The length of the minor axis of the ellipse that fits O by least square method.		[8, 13, 16, 23, 24, 35, 37, 38, 38]
Convex hull perimeter	The perimeter of convex hull $\tilde{O}$ of the shape that enclosed the object O.	$|\partial\tilde{O}|$	[8, 24, 38]
Convex hull area	The area of convex hull.	$|\tilde{O}|$	[8, 13, 16, 17, 20, 24, 37, 38]
Aspect ratio	Ratio between major axis and minor axis.	$\frac{MA}{mA}$	[8, 14, 16, 23, 24, 38, 39]
Eccentricity	Ratio between minor axis and major axis.	$\frac{mA}{MA}$	[8, 14, 23, 24, 38, 39]
Circularity (also know as Form Factor, or compactness)	Measure of how closely the object approaches a mathematically perfect circle (isoperimetric ratio). More sensitive to small scale changes that increase perimeter that roundness.	$\frac{4 \cdot \pi \cdot |O|}{|C|^2}$	[8, 9, 12–14, 16–18, 20, 22–24, 34, 38, 40]

Roundness	As particle becomes circular, value approaches one. Less sensitive to perimeter changes than circularity.	$\dfrac{4 \cdot \lvert O \rvert}{\pi \cdot MA^2}$	[8, 13, 38]
Area equivalent diameter	The diameter of a sphere having the same projection area as particle, or diameter of the circle whose area is same as the cell area.	$\sqrt{\dfrac{4 \cdot \lvert O \rvert}{\pi}}$	[8, 13, 16, 17, 38]
Perimeter equivalent Diameter	The diameter of a circle having the same perimeter as the projection area of the shape.	$\dfrac{\lvert C \rvert}{\pi}$	[8, 13, 17, 38]
Equivalent Ellipse Area	Area of the ellipse that fits O by least square method.	$(\pi \cdot MA \cdot mA)/4$	[8, 24, 38]
Compactness	Degree to which a shape is compact with regard to perfect circle.	$\dfrac{\sqrt{\frac{4 \cdot \lvert O \rvert}{\pi}}}{MA}$	[8, 38, 41]
Solidity	Measures the density of an object. The ratio of the shape area to the area of its convex hull. The proportion of pixels of convex hull that are also in the shape.	$\dfrac{\lvert O \rvert}{\lvert \tilde{O} \rvert}$	[8, 13, 16, 17, 20, 24, 38]
Concavity	Number of pixels that belongs to the convex hull but not to shape. A low concavity implies that the shape is really convex, but in our context this implies that the particle is well rounded.	$\lvert \tilde{O} \rvert - \lvert O \rvert$	[8, 13, 37, 38] [41]
Convexity	A measure of convexity that is less sensitive to area changes. Is a measure of the particle edge roughness.	$\dfrac{\lvert \partial \tilde{O} \rvert}{\lvert C \rvert}$	[8, 12, 38]

Shape	Ratio of squared perimeter to area of the object. The larger its value, the more the shape of the convexity moves away, and it can have holes.	$\frac{|C|^2}{|O|}$	[8, 11–13, 18, 19, 37, 38]
RFactor	Ratio between a circumference of radius $\frac{MA}{2}$ and the booundary of the convex hull	$\frac{|\partial \tilde{O}|}{MA \cdot \pi}$	[8, 15, 38]
Modification ratio	Ratio between the diameter of enclosed circle and the length of the major axis of the ellipse that fits O.	$\frac{2 \cdot mR}{MA}$	[8, 15, 38]
Sphericity	Measures the degree to which a shape approaches the shape of a "sphere".	$\frac{mR}{MR}$	[8, 15, 38]
Bounding box area (BBA)	Area of the rectangle that circumscribes the object.	$MFD \cdot mFD$	[8, 37, 38]
Rectangularity	Ratio of area of the object to its bounding box area.	$\frac{|O|}{\text{BBA}}$	[8, 16, 38]
Shape Factor 1	The ratio of Feret minimun to Feret maximum	$\frac{mFD}{MFD}$	[8, 12, 24, 36, 38]
Elongation	Ratio between Max feret diameter and Min feret diameter.	$\frac{MFD}{mFD}$	[12, 14, 16–18, 24]
Fourier descriptors (FD1,…, FD4)	The magnitudes of the first 4 Fourier descriptors.		[13, 15, 16, 22, 39]

Invariant moments (HU1,..., HU7)	Hu moments.	$I_1 = \eta_{20} + \eta_{02}$	[8, 13, 16, 19, 38]

$$I_1 = \eta_{20} + \eta_{02}$$

$$I_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2$$

$$I_3 = (\eta_{30} - \eta_{12})^2 + (3\eta_{21} - \eta_{03})^2$$

$$I_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2$$

$$I_5 = (\eta_{30} + 3\eta_{12})[(\eta_{30} + \eta_{12})^2$$

$$-3(\eta_{21} + \eta_{03})^2] + (3\eta_{21} - \eta_{03})$$

$$(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2$$

$$-(\eta_{21} + \eta_{03})^2]$$

$$I_6 = (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12}^2)$$

$$-(\eta_{21} + \eta_{03})^2])$$

$$+4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03})$$

$$I_7 = (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})$$

$$[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2]$$

$$-(\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03})$$

$$[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2],$$

where $\eta_{ij} = \frac{\mu_{ij}}{\mu_{00}^{1+\frac{i+j}{2}}}$

and $\mu_{pq} = \sum_x \sum_y (x - \overline{x})^p (y - \overline{y})^q f(x, y)$

where $f(x, y)$ is an image

and $\overline{x} = \frac{M_{10}}{M_{00}}, \overline{y} = \frac{M_{01}}{M_{00}},$

where M_{ij} are raw moments

Table 1: Shape and geometric features extracted from the literature and used in this work for RBC classification.

2.2.2. Color features

Statistical measures based on color spaces are applied to the feature extraction of a cell. The idea is to try to differentiate various cells by color [15]. We used RGB, HSV, and CIEL*a*b* color spaces, as they are reported in the literature to be the most convenient for classification tasks [4, 9, 15, 37, 42]. Let I be a channel of a color space, N be the number of pixels in a region of interest O (each detected cell) in channel I, and $I(p)$ be the intensity value for a pixel p, then the channel color mean μ and its standard deviation σ are respectively calculated as follows:

$$\mu_{I,O} = \frac{1}{N} \sum_{p \in O} I(p), \quad \sigma_{I,O} = \sqrt{\frac{1}{N-1} \sum_{p \in O} (I(p) - \mu)^2}.$$

For each detected cell and channel, we calculated the mean value and the standard deviation, therefore we extracted eighteen color features.

2.2.3. Texture features

From each cell we extracted histogram and grey-level co-occurrence matrix (GLCM) features. Besides GLCM, other textures features has been used in order to classified RBCs in blood smear images, see by example [4, 12, 15, 18, 37]. Histogram features are first order statistical features used to represent texture of the cell's surface, we extracted two histogram features: skewness and kurtosis. The GLCM is widely used for feature extraction in blood cell classification tasks [7, 9, 15, 35, 43]. The GLCM represents the relationship between neighboring pixel intensities and characterizes the texture of an image [44]. In our case, the GLCM was calculated for three pixel distances (1, 2, and 3) and four angles (0, $\pi/4$, $\pi/2$, and $3\pi/4$) for a total of 12 values for each of five GLCM features listed in Table 2, where we defined $P_{i,j}$ as the i,j-th element of the GLCM and N_g as the number of quantised grey tones. Therefore, we extracted 60 GLCM texture features and 62 texture features in total.

2.3. RBC classification

In this section, we present the methods used for RBC classification. First, we explain the problems that we encountered during the data pre-processing and discuss the methods used for solving them. Second, we describe in detail the classifiers we chose to compare. Finally, we describe the process to select the best parameters for each classifier in order to compare them.

2.3.1. Data pre-processing

Before applying classification methods, we needed to analyze the data to resolve potential problems and make classification as accurate as possible [10].

The first problem we encountered was that dataset values resulting from the feature extraction process had different units and ranges. This problem can affect classification results, especially regarding classifiers such as the SVM, MLP, and can also affect feature reduction methods such as principal component analysis (PCA). Non standardized features

Texture feature	Description	Equation
Skewness	Measures the degree of asymmetry.	$\frac{1}{N}\sum_{i=1}^{N}(\frac{x_i-\mu}{\sigma})^3$
Kurtosis	Measure of whether the data are peaked or flat relative to a normal distribution.	$\frac{1}{N}\sum_{i=1}^{N}(\frac{x_i-\mu}{\sigma})^4-3$
Contrast	Measures of the intensity contrast between a pixel a its neighbor pixels. Represents the amount of local gray level variation.	$\sum_{i,j=0}^{N_g-1}P_{i,j}(i-j)^2$
Dissimilarity	Belongs to contrast group of features and its weights are linear. Is a measure of distance between pairs of objects in the region of interest.	$\sum_{i,j=0}^{N_g-1}P_{i,j}|i-j|$
Homogeneity	Measures the smoothness (homogeneity) of the gray level distribution; it is (approximately) inversely correlated with contrast (if contrast is small, usually homogeneity is large).	$\sum_{i,j=0}^{N_g-1}\frac{P_{i,j}}{1+(i-j)^2}$
Energy	Measures the textural uniformity of the gray level distribution and detects disorders in textures. A smaller number of gray levels implies a higher energy.	$\sqrt{\sum_{i,j=0}^{N_g-1}P_{i,j}^2}$
Correlation	Measures the linear dependency of gray levels on those of neighboring pixels.	$\sum_{i,j=0}^{N_g-1}P_{i,j}\frac{(i-\mu_i)(j-\mu_j)}{\sigma_i\cdot\sigma_j}$

Table 2: Texture features extracted from the literature and used in this work for RBC classification. The first two rows are histogram features, and the rest rows are GLCM features.

lead to poor performance due to faster updates of certain weights. To resolve the problem, we used the standard scaling method. We chose this method because it scales the feature values in way that their distribution is centered around 0 with a standard deviation of 1, and because all features are numerical values. Standard scores are calculated as follows:

$$z = \frac{x-\mu}{\sigma}, \tag{1}$$

where μ is the mean and σ is the standard deviation.

2.3.2. Classification

From earlier studies in the area of blood cell classification, we recognized seven different classifiers reported to perform well at a given task. The descriptions of the classifiers are as follows:

- SVM: This statistical learning classifier, in regard to the defined classes, creates a hyper-plane to maximize the separation of data [5, 45]. Different studies proved robust and efficient in blood cell classification tasks [15, 37, 46, 47].

- Decision trees (DT): Non-parametric supervised learning method that uses simple decision rules concluded from data to predict output values [10]. We used the classification and regression trees (CART) algorithm. The main reason that we chose this classifier is its simplicity. Also, Maity et al. [38] reported that DT was very successful in classifying different types of erythrocytes. Furthermore, we wanted to evaluate how this simple method compares to other, more complex algorithms.

- Random forest (RF): It combines a number of fully grown decision tree classifiers developed from various dataset subsamples. It uses averaging to improve the predictive accuracy and to control overfitting. We chose to compare it with other classifiers owing to the evidence that it can outperform SVM [46].

- Extra trees (ET): This classifier is also known as extremely randomized trees [48]. It is a variant of RF and it differs in that it splits nodes by choosing random cut-points and uses the whole learning sample at each step. In [49] satisfactory performance of ET was reported on an RBC dataset of isolated cells, taken in a capillary 0.05×0.5 mm suspension of cells using an ACM microscope [50].

- Gradient boosting (GB): It combines weak classifiers, same as RF, and reduces the resulting error. It combines weak decision trees, which contain two leaves and have high bias and low variance. GB reduces the error by lowering the bias [10].

- k-nearest neighbors (kNN): This classification method is a type of non-generalized learning. It does not create a general model, but instead stores instances of the training data. Classification is computed from a majority vote of the nearest neighbors of each point. We used it because it has performed well when classes have irregular boundaries, although it is a simple classifier [10]. In [16] 33 features were used as inputs to three machine learning classification methods: Neural Networks, SVM and kNN for RBC classification by using maximum voting.

- Multilayer perceptron (MLP): It is the one of the most widely used artificial neural networks algorithms. It is composed of layers of input and output elements related through weighted connections. We used a MLP algorithm that learns using backpropagation. It has been reported as performing well for blood cell classification in general [42], and we wanted to test it for classification of red blood cells exclusively.

2.3.3. Hyperparameter Selection

Once we selected the classifiers, we wanted to compare them. First we selected the best parameters for each one, then we compared the performance of each classifier. Selecting the best parameters to compare all classifiers was conducted by two methods: Randomized and Grid search.

Randomized search method randomly generates the subset of a given range of possible parameter values and evaluates the performance regarding the value of the defined metric. Grid search is an exhaustive search over every subset of the defined hyper-parameter space.

To avoid overfitting we used a 10-fold cross-validation procedure [10] to evaluate classifiers performance using both Randomized and Grid search. Dataset was split into training and testing sets with ratio 70/30. Training set was used for fine tuning the classifiers and the prediction was made on the testing set using the best parameters resulting from the Randomized and Grid search.

3. Experimental Framework

In order to validate our work, we presented a case study for microscopic images of patients with SCD. In this section we described in detail the dataset that we used. Furthermore, we explained the metrics used to compare the classifiers. Finally, we described the conducted experiment, which has three parts. First part consists of tuning classifiers to find the best parameters for given dataset with all features. In second part, we used feature reduction and selection techniques to improve the results from the first part of the experiment and also to advance towards interpretable models. Finally, we compared our results with the state of the art of RBC morphological classification. All the cells were classified regarding previously defined three classes.

All the experimentation of this article have been performed using Python 3.6. All classifiers used, belong to the Python scikit-learn library. The code that we implemented is available at `https://gitlab.com/miquelca32/features/`. Experiments were run on the machine with AMD Ryzen 5 3600 6-Core processor 3.60 GHz, 32 GB of RAM and Windows 10 operating system.

3.1. Dataset

The microscopic images of the blood smears used in this work were collected from erythrocytesIDB [34], available at http://erythrocytesidb.uib.es/. The images consist in prepared samples of patients with sickle cell anemia classified by a specialist from "Dr. Juan Bruno Zayas" Hospital General in Santiago de Cuba, Cuba.

Samples were obtained from voluntary donors. The donor's thumb was pricked with a lancet and a drop of blood was collected on a sheet. The blood was spread with a coverslip at an angle of 45° with respect to the sheet, allowed to dry, and then fixed with a May-Grünwald methanol solution. May-Grünwald-Giemsa staining is a frequently used method for blood smears.

The images were acquired using a Leica microscope (100×) and a Kodak EasyShare V803 camera (Kodak Retinar Aspheric All Glass Lens of 36–108 mm AF 3× optical).

Every image was labeled by the medical expert and the specialist's criteria were used as an expert approach to validate the results of the classification methods used in this paper.

In the dataset there are a total of 2695 tagged cells (1663 circular, 700 elongated and 332 other cells). Overlapping cells were included with a total of 66 clusters (7.1% of the cells).

3.2. Evaluation metrics

To be able to compare the performance of different classifiers we used the well-known F-measure, and SDS-score [51] as an indicator that the results provided by the method are useful for the diagnosis support of SCD. SDS-score is justified with the fact that the misclassification of the normal cells as the elongated or other cells will cause the alert to the medical specialist that the patient's condition has worsened and that the therapy should be changed. Then, it is up to the specialist to review the diagnosis and to decide whether the more drastic treatment should be prescribed. This type of error is not so serious because the treatment usually has no side effects. More dangerous scenario would be to classify deformed cells (elongated or other) as normal. In this case, the specialist could decide that the patient is not at risk of a vaso-occlusive crisis, and the necessary treatment would not be applied. To support the diagnosis in a good way, classifiers needs to minimize the missclassification rate of elongated cells and cells with other deformations as normal cells, and the missclassification of normal cells as elongated and cells with other deformations.

We also used Class Balance Accuracy (CBA) [52] and Mathews Correlation Coefficient (MCC) [53] for imbalanced classes, due to the use of overall field of view of the image, where the existing quantity of normal cells is much greater than the quantity of elongated or other deformation cells. The study of performance measures for multiple class problems is an active research field [54, 55].

All of these measures are derived from the confusion matrix of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) for all three classes.

3.3. Experiment 1: Selection of best classifiers

The goal of the first experiment was to select the best performing classifiers on the available dataset. The best classifiers are the result of the performance comparison described in Section 3.2.

First, we calculated each classifier performance using the baseline parameters defined by the Python's Scikit library and over all the set of features using 10-fold cross-validation procedure to avoid overfitting. Second, we performed a fine tuning procedure on each of them. Dataset was split to train and test set in 70/30 ratio. The fine tuning consisted of selecting the values for each parameter that ensures the best performance of each classifier regarding the average F-measure, and then regarding the SDS-score. Fine tuning was performed using randomized search with 10-fold cross-validation on the training set. Having selected the best parameters for every classifier, we made a prediction on the testing set to obtain the final results. We compared all the results and finally, selected the best performing classifiers.

17

3.4. Experiment 2: Selection of the most important features

The goal of the second experiment was to select the most important features for this classification task while preserving or even improving the performance of the chosen classifiers. The idea was to reduce the number of the features because they may cause unnecessary noise in the classification process. Furthermore, the reduction of features decreases the model's complexity and training time. Also, as features are an important component of interpretable machine learning [31] we would like to rank them by their importance to analyse if this can explain the diagnosis.The feature selection was conducted in two parts.

First, we took best classifiers from previous experiment and we ranked their features by their importance. Then we trained each classifier using a wrapper method to select the best feature set, starting from the best one and iteratively increasing the number of features used, until the classifier performance was maximized. At each iteration best parameters were found performing a grid search with 10-fold cross-validation on the training set. The evaluation metrics were the F-measure and the SDS-score. The results were obtained using the best parameters to make prediction on the testing set.

Second, we tested feature reduction techniques, that is, Principal Component Analysis (PCA) and Linear discriminant analysis (LDA). Regarding PCA, we tested best classifiers from first experiment with different number of principal components, up to explain more than 95% of variance.

Finally we wanted to compare the different sets of best features we obtained. We tested if this sets could improve the results obtained in first experiment for all classifiers. We conducted the fine tuning performing a grid search with 10-fold cross-validation on the training set of each classifier described in Section 2.3.3 with the different sets of features obtained from PCA, LDA and the ones obtained from best classifiers.

3.5. Experiment 3: Comparison with the state-of-art

The goal of the third experiment is to compare best classifiers selected from previous experiment with other state-of-art algorithms. The follow state-of-the-art on automatic counter of RBCs based on erythrocyte shape descriptors were considered in order to compare them with the method proposed in this paper.

In [40], Asakura determines the percentage of sickle erythrocytes cells with three different morphologies by measuring the area, perimeter, and short-axis/long-axis ratio of each cell, that is, the CSF (circular shape factor) and ESF (elliptical shape factor), that are our circularity and eccentricity, respectively (see Table 1).

In [56], Acharya and Kumar presented a method for automatically classify RBCs in two classes: normal cell (circular shape) and abnormal cell. After a conversion to grey scale, the image segmentation was performed using the global Otsu threshold method. Morphological operations and modified watershed transform were applied to the image segmentation to separate the touching objects. After that, two possibilities are open: counting the cells using the labeling algorithm and the CSF (form factor or our circularity) to identify normal cells, or counting the cells using circular Hough transform.

Classifier	SDS-score Parameters	F-measure Parameters
SVM	kernel='rbf', C=10, gamma=0.01	kernel='rbf', C=10, gamma=0.01
DT	max_depth=None, max_features=None, min_samples_leaf=10, min_samples_split=10, splitter='best'	max_depth=None, max_features=None, min_samples_leaf=10, min_samples_split=10, splitter='best'
RF	n_estimators=300, min_samples_leaf=2, min_samples_split=2, max_depth=None, max_features='sqrt', bootstrap=True	n_estimators=300, min_samples_leaf=2, min_samples_split=2, max_depth=None, max_features='sqrt', bootstrap=True
ET	n_estimators=60, max_depth=None, max_features='auto', min_samples_leaf=1, min_samples_split=10, bootstrap=True	n_estimators=60, max_depth=None, max_features='auto', min_samples_leaf=1, min_samples_split=10, bootstrap=True
GB	sub_sample=1, min_child_weight=1, max_depth=10, max_delta_step=10	sub_sample=1, min_child_weight=1, max_depth=10, max_delta_step=20
kNN	n_neighbors=10, weights='distance', p=1, algorithm='brute'	n_neighbors=9, weights='distance', p=1, algorithm='brute'
MLP	solver='sgd', hidden_layer_sizes=(10, 3), activation='identity'	solver='sgd', hidden_layer_sizes=(10, 3), activation='identity'

Table 3: Best parameters selected by fine tuning that maximize the SDS-score and F-measure performance measures,. respectively.

In [34], González-Hidalgo *et al.* proposed a method for the analysis of the shape of erythrocytes in peripheral blood smear samples of sickle cell disease, which uses ellipse adjustments and an algorithm for detecting notable points. They applied a set of constraints that allowed to eliminate significant image preprocessing steps proposed in previous studies. They classify cells between circular and elongated using the ESF (our eccentricity).

4. Results and discussion

The aims of the experiments are to determine the most appropriate classifier to successfully support the diagnose of the SCD, to find the set of most important features and to compare our results with other state-of-the-art algorithms.

4.1. Experiment 1

Recall that, as was pointed in Section 3.3, this experiment was designed to select best classifiers for the cell classification task. Table 3 shows best parameters for each classifier regarding the used evaluation measures.

Table 4 depicts the comparison of classifiers regarding F-measure and SDS-score. From the obtained results we can observe that GB, RF, DT, ET and MLP have similar performance, although the best ones are RF and GB. Regarding running times of randomized search over the same dataset, we excluded GB. It took 132 minutes for GB, 11 minutes for RF, 7 minutes for ET and only 50s for DT to run. Therefore, we chose RF as the best resulting classifier and DT as second best and fastest one to proceed with the experiments (taking account SDS-score results because indicates how useful are the results provided by the method for support the diagnosis of SCD). Moreover, Appendix C shows the running times of the experimentation, where running time of GB, for one iteration of Grid search and for training, takes much longer than the rest of the classifiers. The benefits obtained by classification using GB is

not compensated by the time needed to train this classifier and perform the fine-tuning process. Confusion matrices obtained from best classifiers as a result of this process are depicted in Table 5 (Appendix A shows confusion matrices for each classifier).

	SVM	DT	RF	ET	GB	kNN	MLP
Baseline F-measure	85.97%	86.49%	89.91%	89.54%	89.75%	85.35%	85.39%
Fine tuning F-measure	91.50%	91.62%	**93.11%**	92.91%	**93.40%**	88.93%	92.23%
Baseline SDS-score	87.24%	90.02%	91.61%	89.76%	90.91%	87.31%	87.68%
Fine tuning SDS-score	93.45%	94.44%	**94.56%**	94.31%	**94.56%**	90.98%	94.19%

Table 4: Performance of the different classifiers for the F-measure and SDS-score performance measures in Experiment 1. We show the performance measures for the baseline experiment and the fine tuning parameters experiment.

RF baseline	c	e	o
c	1580	38	45
e	52	636	12
o	91	27	214

RF max F-measure	c	e	o
c	490	7	2
e	9	197	4
o	26	6	68

RF max SDS-score	c	e	o
c	490	7	2
e	9	197	4
o	26	6	68

DT baseline	c	e	o
c	1535	47	81
e	69	564	67
o	72	32	228

DT max F-measure	c	e	o
c	481	8	10
e	6	194	10
o	21	12	67

DT max SDS-score	c	e	o
c	481	8	10
e	6	194	10
o	21	12	67

Table 5: Confusion matrices for best classifiers selected in the baseline experiment (left) and fine tuning to maximizing F-measure and SDS-score (center and right).

4.2. Experiment 2

First part of the experiment consisted on using best classifiers from previous experiment to check if there were a subset of features that improve the previous obtained results. We performed an incremental search over the ordered set of all features, for RF and GB, based on its importance. We ranked the RF features based on its mean decrease

impurity and we ranked the GB features based on the average of the impurity gain across each of the classifiers. For each subset of ordered features we performed a grid search of the best parameters in order to maximize the SDS-score and the F-measure. Second column of Table 6 shows that we improved the results obtained in the first experiment, we can observe that dimensionality reduction gives us an improvement over results of first experiment. RF result is maximized (SDS-score 95.06%) with the 15 most important features: Aspect ratio, Elongation, R factor, HU1, Roundness, HU2, Minor axis, Shape, FD1, Circularity, Min feret, HU7, HU3, Compactness, and Skewness. Although DT result is maximized (SDS-score 94.44%) with all the features, we considered most important features to be the first 20, when the result was over 94% (SDS-score 94.07%) and F-measure maximized with 92.37%: R factor, Aspect ratio, Area equivalent diameter, Skewness, Roundness, Minor axis, Correlation1, Homogeneity12, Max R, Major axis, HU5, Min feret, Energy9, Homogeneity1, Solidity, Kurtosis, Blue mean, HU3, Contrast9, L mean.

Second part of this experiment was to reduce the feature space using two dimensionality reduction methods, PCA and LDA over the whole set of features. We performed a search on the PCA dimensional space in order to maximize the classification measures up to explain the 95.01% of the variance, that is selecting the first 16 components. For LDA we used 2 components explaining 100% of variance. First component explained 83.97% of variance. For each subset of ordered principal components we performed a grid search on RF and DT. In Table 6, we can state that LDA dimensional reduction outperforms PCA, but the feature selection outperforms both PCA and LDA. More precisely, in our case study reducing features using PCA and LDA worsened the results of the first experiment.

	First experiment	Feature selection	PCA	LDA
RF F-measure	93.11%	**93.36**%	90.06%	91.55%
RF SDS-score	94.56%	**95.05**%	92.34%	93.20%
DT F-measure	91.62%	92.37%	87.02%	91.49%
DT SDS-score	94.44%	94.07%	90.48%	93.20%

Table 6: Comparison of the performance measures obtained using fine tuning in Experiment 1, with the obtained feature selection, PCA and LDA dimensionality reduction. For RF and DT machine learning methods we maximized the SDS-score and the F-measure.

Table 7 depicts the comparison between results from the first experiment and classification with 15 most important features regarding RF on all the classifiers. Results from DT are also obtained using RF's most important features, as they are better than using DT's most important features as showed in Table 6. We can observe that feature selection regarding RF and the posterior fine tuning of the parameters of each classifier gives the best performance for the SDS-score and F-measure. We can note that method with best results is GB for both measures but the results for RF are second best, for that reason these are the two selected machine leaning methods for the third experiment. Appendix B shows confusion matrices for each classifier.

Regarding interpretability, we can observe in Table 7 that metrics for classifiers are not so different. Their ranges

	SVM	DT	RF	ET	GB	kNN	MLP
Baseline SDS-score	87.24%	90.02%	91.61%	89.76%	90.91%	87.31%	87.68%
Important features SDS-score	94.68%	94.68%	**95.05%**	94.31%	**95.18%**	93.57%	94.81%
Baseline F-measure	85.97%	86.49%	89.91%	89.54%	89.75%	85.35%	85.39%
Important features F-measure	93.15%	92.55%	**93.36%**	92.88%	**93.50%**	91.77%	93.14%

Table 7: Comparison of the performance measures between the obtained in baseline, Experiment 1, and the obtained by the different machine learning methods using feature selection.

are from 93.57% to 95.18% for the SDS-Score, and from 91.77% to 93.50% for the F-measure. DT and kNN are transparent models [57], that means that the model by itself is understandable. Moreover, we ranked the features as a first step towards post-hoc interpretability [57] to target models that are not readily interpretable by design by resorting to diverse means to enhance their interpretability such as feature relevance. As can be observed, in both machine learning methods, DT and RF, only shape/geometric and texture features were selected and leaving it aside color features. This makes sense since we are classifying shapes according to their morphology, and color has no influence on it. Feature relevance explanation techniques aim to describe the functioning of an opaque model by ranking relevance or importance each feature has in the prediction output by the model to be interpret.

	GB	RF	[40]
SDS-score	**95.18%**	95.05%	61.80%
F-measure	**93.50%**	93.36%	45,33%
CBA	**88.39%**	88.06%	37.48%
MCC	**88.43%**	88.20%	35.43%

Table 8: Comparison of F-measure and SDS-score values obtained by GB, RF and Asakura [40] methods for the 3 classes classification problem.

4.3. Experiment 3

As was stated in Section 3.5, the aim of this experiment is to compare our results with other state-of-the-art algorithms. First, we compared with the method described by Asakura *et al.* in [40] as they classify eritrocites in 3 classes. Table 8 shows that our two best classifiers with F-measure and SDS-score (GB and RF) outperform results from Asakura [40]. Table 9 shows the confusion matrices used to obtain the metrics of this part of the experiment.

The second part of this validation experiment was to compare with other methods that classify cells in two classes: circular and elongated. In particular we compare with the work of Acharya *et al.* [56] and González-Hidalgo *et al.* [34]. Table 10 shows that our two best classifiers with F-measure and SDS-score (GB and RF) outperform results

from Acharya [56] and González-Hidalgo [34]. Table 11 shows the confusion matrices used to obtain the metrics of this part of the experiment.

GB	c	e	o	RF	c	e	o	[40]	c	e	o
c	488	6	5	c	488	7	4	c	199	259	41
e	8	194	8	e	7	194	9	e	0	102	108
o	20	5	75	o	22	4	74	o	9	67	24

Table 9: Confusion matrices for best classifiers (GB and RF) and that obtained by Asakura [40].

	GB	RF	[56]	[34]
SDS-score	**94.68%**	94.44%	78.49%	49.32%
F-measure	**94.67%**	94.42%	78.76%	48.97%
CBA	**93.98%**	93.66%	81.16%	52.81%
MCC	**88.72%**	88.19%	60.80%	5.70%

Table 10: Comparison of F-measure and SDS-score values obtained by GB, RF, Acharya [56] and González-Hidalgo [34] methods for the 2 classes classification problem.

GB SDS-score	c	e	RF SDS-score	c	e	[34]	c	e	[56]	c	e
c	484	15	c	484	15	c	189	310	c	348	151
e	28	282	e	30	280	e	100	210	e	23	287

Table 11: Confusion matrices for GB, RF and results from Acharya [56] and González-Hidalgo [34].

Regarding interpretability, we can observe in Table 7 that transparent models [57], DT and kNN, obtained better results than [34, 40, 56].

5. Conclusion and future work

We presented a general methodology to select the classification method and features with best performance for diagnostic support through peripheral blood smear images of red blood cells, in our case samples of patients with sickle-cell disease diagnosis which can be generalized for other study cases.

We collected and implemented features and classifiers that previous literature validated for the study case or similar cases. Then, we explained how to select the best parameters for each classifier. We included as results the best parameters for each classifier, the implemented code `https://gitlab.com/miquelca32/features/` and the confusion matrices with the raw data to allow researchers to more easily compute other metrics. The dataset we used is available at `http://erythrocytesidb.uib.es/`. For the sake of scientific progress, it would be beneficial if authors published their raw data, code and the image datasets that they used.

Next, we defined how to select the most important features for classification in order to reduce their total number to decrease the complexity and training time. Also, feature selection allowed us to give more information about the functioning of opaque models (non interpretable) by feature ranking in the model's prediction output.

We compared the best performing classification methods with the state-of-the-art methods in order to validate that the proposed methodology allows to select the most appropriate ML model for a case study, in this paper microscopic images of patients with sickle cell anemia. More precisely, we obtained better results than the state-of-the-art methods with all the models used in this paper, even with interpretable model classifiers. That is a good result for health environments where there is a need to trust and audit diagnostic support systems.

As a further work, we want to research on diagnosis support of other diseases through cell morphology analysis from microscopy images and advance towards more interpretable approaches. We also are interested in research on ensemble voting, subset of features for optimizing classification, and feature selection based on PCA and LDA for sickle-cell disease diagnosis support and its interpretability.

References

References

[1] W. H. Organization, Sickle-cell disease and other haemoglobin disorders, `https://healthservices.uonbi.ac.ke/sites/default/files/centraladmin/healthservices/Sickle-cell%20disease%20and%20other%20haemoglobin%20disorders.pdf`, ac-

cessed: 2019-10-23 (2011).

[2] I. Abubakar, T. Tillmann, A. Banerjee, Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the global burden of disease study 2013, Lancet 385 (9963) (2015) 117–171.

[3] M. Naghavi, A. A. Abajobir, C. Abbafati, K. M. Abbas, F. Abd-Allah, S. F. Abera, V. Aboyans, O. Adetokunboh, A. Afshin, A. Agrawal, et al., Global, regional, and national age-sex specific mortality for 264 causes of death, 1980–2016: a systematic analysis for the global burden of disease study 2016, The Lancet 390 (10100) (2017) 1151–1210.

[4] A. Merino, L. Puigví, L. Boldú, S. Alférez, J. Rodellar, Optimizing morphology through blood cell image analysis, International journal of laboratory hematology 40 (2018) 54–61.

[5] C. Zhu, B. Zou, R. Zhao, J. Cui, X. Duan, Z. Chen, Y. Liang, Retinal vessel segmentation in colour fundus images using extreme learning machine, Computerized Medical Imaging and Graphics 55 (2017) 68–77.

[6] J. Shi, S. Zhou, X. Liu, Q. Zhang, M. Lu, T. Wang, Stacked deep polynomial network based representation learning for tumor classification with small ultrasound image dataset, Neurocomputing 194 (2016) 87–94.

[7] D. K. Das, M. Ghosh, M. Pal, A. K. Maiti, C. Chakraborty, Machine learning approach for automated screening of malaria parasite using light microscopic images, Micron 45 (2013) 97–106.

[8] M. Maity, T. Mungle, D. Dhane, A. K. Maiti, C. Chakraborty, An ensemble rule learning approach for automated morphological classification of erythrocytes, Journal of medical systems 41 (4) (2017) 56.

[9] J. Rodellar, S. Alférez, A. Acevedo, A. Molina, A. Merino, Image processing and machine learning in the morphological analysis of blood cells, International Journal of Laboratory Hematology 40 (S1) (2018) 46–53. `arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/ijlh.12818`, `doi:10.1111/ijlh.12818`.
URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/ijlh.12818`

[10] J. Friedman, T. Hastie, R. Tibshirani, The elements of statistical learning, Springer Series in Statistics, Springer New York Inc., New York, NY, USA, 2001.

[11] J. W. Bacus, M. G. Belanger, R. K. Aggarwal, J. F E Trobaugh, Image processing for automated erythrocyte classification., Journal of Histochemistry & Cytochemistry 24 (1) (1976) 195–201, pMID: 1254916. `arXiv:https://doi.org/10.1177/24.1.1254916`, `doi:10.1177/24.1.1254916`.
URL `https://doi.org/10.1177/24.1.1254916`

[12] L. L. Wheeless, R. D. Robinson, O. P. Lapets, C. Cox, A. Rubio, M. Weintraub, L. J. Benjamin, Classification of red blood cells as normal, sickle, or other abnormal, using a single image analysis feature, Cytometry 17 (2) (1994) 159–166. `arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/cyto.990170208`, `doi:10.1002/cyto.990170208`.
URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/cyto.990170208`

[13] D. DAS, C. CHAKRABORTY, B. MITRA, A. MAITI, A. RAY, Quantitative microscopy approach for shape-based erythrocytes characterization in anaemia, Journal of Microscopy 249 (2) (2013) 136–149. `arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/jmi.12002`, `doi:10.1111/jmi.12002`.
URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/jmi.12002`

[14] V. Acharya, P. Kumar, Identification and red blood cell classification using computer aided system to diagnose blood disorders, in: 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2017, pp. 2098–2104. `doi:10.1109/ICACCI.2017.8126155`.

[15] J. A. Akrimi, A. Suliman, L. E. George, A. R. Ahmad, Classification red blood cells using support vector machine, in: Information Technology and Multimedia (ICIMU), 2014 International Conference on, IEEE, 2014, pp. 265–269.

[16] M. Lotfi, B. Nazari, S. Sadri, N. K. Sichani, The detection of dacrocyte, schistocyte and elliptocyte cells in iron deficiency anemia, in: Pattern Recognition and Image Analysis (IPRIA), 2015 2nd International Conference on, IEEE, 2015, pp. 1–5.

[17] L. F. Rodrigues, M. C. Naldi, J. F. Maari, Morphological analysis and classification of erythrocytes in microscopy images, in: Proceedings of XII Workshop de Visão Computacional, Vol. 1, UCDB, 2016, pp. 69–74.

25

[18] M. Veluchamy, K. Perumal, T. Ponuchamy, Feature extraction and classification of blood cells using artificial neural network, American journal of applied sciences 9 (5) (2012) 615.

[19] R. Tomari, W. N. W. Zakaria, M. M. A. Jamil, F. M. Nor, N. F. N. Fuad, Computer aided system for red blood cell classification in blood smear image, Procedia Computer Science 42 (2014) 206–213.

[20] H. A. Elsalamony, Detection of some anaemia types in human blood smears using neural networks, Measurement Science and Technology 27 (8) (2016) 085401.
URL http://stacks.iop.org/0957-0233/27/i=8/a=085401

[21] H.-M. Chen, Y.-T. Tsao, S.-N. Tsai, Automatic image segmentation and classification based on direction texton technique for hemolytic anemia in thin blood smears, Machine vision and applications 25 (2) (2014) 501–510.

[22] X. Gual-Arnau, S. Herold-García, A. Simó, Erythrocyte shape classification using integral-geometry-based methods, Med. Biol. Engineering and Computing 53 (7) (2015) 623–633. doi:10.1007/s11517-015-1267-x.
URL https://doi.org/10.1007/s11517-015-1267-x

[23] V. Sharma, A. Rathore, G. Vyas, Detection of sickle cell anaemia and thalassaemia causing abnormalities in thin smear of human blood sample using image processing, in: 2016 International Conference on Inventive Computation Technologies (ICICT), Vol. 3, 2016, pp. 1–5. doi:10.1109/INVENTIVE.2016.7830136.

[24] M. Xu, D. P. Papageorgiou, S. Z. Abidi, M. Dao, H. Zhao, G. E. Karniadakis, A deep convolutional neural network for classification of red blood cells in sickle cell anemia, PLOS Computational Biology 13 (10) (2017) 1–27. doi:10.1371/journal.pcbi.1005746.
URL https://doi.org/10.1371/journal.pcbi.1005746

[25] M. Zhang, X. Li, M. Xu, Q. Li, Rbc semantic segmentation for sickle cell disease based on deformable u-net, in: A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, G. Fichtinger (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, Springer International Publishing, Cham, 2018, pp. 695–702.

[26] A. Kihm, L. Kaestner, C. Wagner, S. Quint, Classification of red blood cell shapes in flow using outlier tolerant machine learning, PLOS Computational Biology 14 (6) (2018) 1–15. doi:10.1371/journal.pcbi.1006278.
URL https://doi.org/10.1371/journal.pcbi.1006278

[27] L. Alzubaidi, O. Al-Shamma, M. A. Fadhel, L. Farhan, J. Zhang, Classification of red blood cells in sickle cell anemia using deep convolutional neural network, in: A. Abraham, A. K. Cherukuri, P. Melin, N. Gandhi (Eds.), Intelligent Systems Design and Applications, Springer International Publishing, Cham, 2020, pp. 550–559.

[28] L. Alzubaidi, M. A. Fadhel, O. Al-Shamma, J. Zhang, Y. Duan, Deep learning models for classification of red blood cells in microscopy images to aid in sickle cell anemia diagnosis, Electronics 9 (3) (2020) 427.

[29] S. S. Devi, J. Singha, M. Sharma, R. H. Laskar, Erythrocyte segmentation for quantification in microscopic images of thin blood smears, Journal of Intelligent & Fuzzy Systems 32 (4) (2017) 2847–2856.

[30] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (xai), IEEE Access 6 (2018) 52138–52160.

[31] Z. C. Lipton, The mythos of model interpretability, Queue 16 (3) (2018) 31–57.

[32] S. Savkare, S. Narote, Automatic system for classification of erythrocytes infected with malaria and identification of parasite's life stage, Procedia Technology 6 (2012) 405–410.

[33] J. Canny, A computational approach to edge detection, IEEE Transactions on pattern analysis and machine intelligence PAMI-8 (6) (1986) 679–698.

[34] M. González-Hidalgo, F. Guerrero-Pena, S. Herold-García, A. Jaume-i-Capó, P. D. Marrero-Fernández, Red blood cell cluster separation from digital images for use in sickle cell disease, IEEE journal of biomedical and health informatics 19 (4) (2015) 1514–1525.

[35] S. Bhowmick, D. K. Das, A. K. Maiti, C. Chakraborty, Structural and textural classification of erythrocytes in anaemic cases: a scanning electron microscopic study, Micron 44 (2013) 384–394.

[36] M. Yang, K. Kpalma, J. Ronsin, A survey of shape feature extraction techniques (2008).

[37] T. Markiewicz, S. Osowski, B. Marianska, L. Moszczynski, Automatic recognition of the blood cells of myelogenous leukemia using svm,

in: Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on, Vol. 4, IEEE, 2005, pp. 2496–2501.

[38] M. Maity, P. Sarkar, C. Chakraborty, Computer-assisted approach to anemic erythrocyte classification using blood pathological information, in: Emerging Applications of Information Technology (EAIT), 2012 Third International Conference on, IEEE, 2012, pp. 116–121.

[39] N. Deb, S. Chakraborty, A noble technique for detecting anemia through classification of red blood cells in blood smear, in: International Conference on Recent Advances and Innovations in Engineering (ICRAIE-2014), 2014, pp. 1–9. `doi:10.1109/ICRAIE.2014.6909137`.

[40] T. Asakura, T. Hirota, A. T. Nelson, M. P. Reilly, K. Ohene-Frempong, Percentage of reversibly and irreversibly sickled cells are altered by the method of blood drawing and storage conditions, Blood Cells, Molecules, and Diseases 22 (3) (1996) 297–306.

[41] C. Chang, W. Liu, H. Zhang, Image retrieval based on region shape similarity, in: M. M. Yeung, C.-S. Li, R. W. Lienhart (Eds.), Storage and Retrieval for Media Databases 2001, Vol. 4315, International Society for Optics and Photonics, SPIE, 2001, pp. 31 – 38. `doi:10.1117/12.410947`.
URL `https://doi.org/10.1117/12.410947`

[42] G. Ongun, U. Halici, K. Leblebicioglu, V. Atalay, M. Beksaç, S. Beksaç, Feature extraction and classification of blood cells for an automated differential blood count system, in: Neural Networks, 2001. Proceedings. IJCNN'01. International Joint Conference on, Vol. 4, IEEE, 2001, pp. 2461–2466.

[43] S. H. Rezatofighi, H. Soltanian-Zadeh, Automatic recognition of five types of white blood cells in peripheral blood, Computerized Medical Imaging and Graphics 35 (4) (2011) 333–343.

[44] R. M. Haralick, K. Shanmugam, et al., Textural features for image classification, IEEE Transactions on systems, man, and cybernetics SMC-3 (6) (1973) 610–621.

[45] N. Cristianini, J. Shawe-Taylor, et al., An introduction to support vector machines and other kernel-based learning methods, Cambridge university press, 2000.

[46] B. Ko, J. Gim, J. Nam, Cell image classification based on ensemble features and random forest, Electronics Letters 47 (11) (2011) 638–639.

[47] S. R. Amendolia, G. Cossu, M. Ganadu, B. Golosio, G. Masala, G. M. Mura, A comparative study of k-nearest neighbour, support vector machine and multi-layer perceptron for thalassemia screening, Chemometrics and Intelligent Laboratory Systems 69 (1-2) (2003) 13–20.

[48] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Machine learning 63 (1) (2006) 3–42.

[49] R. Marée, P. Geurts, L. Wehenkel, Random subwindows and extremely randomized trees for image classification in cell biology, BMC Cell Biology 8 (1) (2007) S2. `doi:10.1186/1471-2121-8-S1-S2`.
URL `https://doi.org/10.1186/1471-2121-8-S1-S2`

[50] M. Schönfeld, R. Grebe, Automatic shape quantification of freely suspended red blood cells by isodensity contour tracing and tangent counting, Computer Methods and Programs in Biomedicine 28 (4) (1989) 217 – 224. `doi:https://doi.org/10.1016/0169-2607(89)90154-5`.
URL `http://www.sciencedirect.com/science/article/pii/0169260789901545`

[51] W. Delgado-Font, M. Escobedo-Nicot, M. González-Hidalgo, S. Herold-Garcia, A. Jaume-i-Capó, A. Mir, Diagnosis support of sickle cell anemia by classifying red blood cell shape in peripheral blood images, Medical & Biological Engineering & Computing online 28 March 2020 (2020) 1–20. `doi:10.1007/s11517-019-02085-9`.

[52] L. Mosley, A balanced approach to the multi-class imbalance problem, Ph.D. thesis, Iowa State University, Industrial and Manufacturing Systems Engineering Department (2013).

[53] J. Gorodkin, Comparing two k-category assignments by a k-category correlation coefficient, Comput. Biol. Chem. 28 (5-6) (2004) 367–374. `doi:10.1016/j.compbiolchem.2004.09.006`.
URL `http://dx.doi.org/10.1016/j.compbiolchem.2004.09.006`

[54] T. Kautz, B. M. Eskofier, C. F. Pasluosta, Generic performance measure for multiclass-classifiers, Pattern Recognition 68 (2017) 111–125.

[55] P. Branco, L. Torgo, R. P. Ribeiro, Relevance-based evaluation metrics for multi-class imbalanced domains, in: J. Kim, K. Shim, L. Cao, J.-G. Lee, X. Lin, Y.-S. Moon (Eds.), Advances in Knowledge Discovery and Data Mining, Springer International Publishing, Cham, 2017, pp. 698–710.

[56] V. Acharya, P. Kumar, Identification and red blood cell automated counting from blood smear images using computer-aided system, Medical & biological engineering & computing 56 (3) (2018) 483–489.

[57] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, Information Fusion 58 (2020) 82–115.

Appendix A. First experiment: confusion matrices

SVM baseline	c	e	o
c	1496	44	123
e	83	596	21
o	94	19	219

SVM max F-measure	c	e	o
c	1559	50	54
e	81	615	4
o	111	31	190

SVM max SDS-score	c	e	o
c	1559	50	54
e	81	615	4
o	111	31	190

DT baseline	c	e	o
c	1535	47	81
e	69	564	67
o	72	32	228

DT max F-measure	c	e	o
c	1552	40	71
e	51	620	29
o	82	32	218

DT max SDS-score	c	e	o
c	1552	40	71
e	51	620	29
o	82	32	218

RF baseline	c	e	o
c	1580	38	45
e	52	636	12
o	91	27	214

RF max F-measure	c	e	o
c	1596	40	27
e	39	657	4
o	103	33	196

RF max SDS-score	c	e	o
c	1597	41	25
e	40	654	6
o	103	32	197

ET baseline	c	e	o
c	1556	45	62
e	72	620	8
o	97	29	206

ET max F-measure	c	e	o
c	1569	47	47
e	54	637	9
o	101	32	199

ET max SDS-score	c	e	o
c	1569	47	47
e	54	637	9
o	101	32	199

GB baseline	c	e	o
c	1569	41	53
e	60	628	12
o	91	26	215

GB max F-measure	c	e	o
c	1565	45	53
e	56	631	13
o	82	24	226

GB max SDS-score	c	e	o
c	1569	41	53
e	59	628	13
o	78	24	230

kNN baseline	c	e	o
c	1542	39	82
e	109	577	14
o	112	31	189

kNN max F-measure	c	e	o
c	1551	43	69
e	111	579	10
o	112	29	191

kNN max SDS-score	c	e	o
c	1551	43	69
e	111	579	10
o	112	29	191

MLP baseline	c	e	o
c	1485	48	130
e	62	598	40
o	92	29	211

MLP max F-measure	c	e	o
c	1515	45	103
e	66	618	16
o	75	25	232

MLP max SDS-score	c	e	o
c	1539	45	79
e	59	598	43
o	82	30	220

Table A.12: Confusion matrices for the first experiment (raw data Table 4).

Appendix B. Second experiment: confusion matrices

RF 15 features SDS-score	c	e	o
c	488	7	4
e	7	194	9
o	22	4	74

RF 15 features F-measure	c	e	o
c	489	7	3
e	8	194	8
o	22	3	75

DT 20 features max SDS-score	c	e	o
c	483	13	3
e	8	199	3
o	24	9	67

DT 20 features F-measure	c	e	o
c	483	13	3
e	8	199	3
o	24	9	67

RF PCA max SDS-score	c	e	o
c	485	9	5
e	14	194	2
o	34	12	54

RF PCA max F-measure	c	e	o
c	485	9	5
e	14	194	2
o	34	12	54

DT PCA max SDS-score	c	e	o
c	483	9	7
e	25	172	13
o	36	11	53

DT PCA max F-measure	c	e	o
c	483	9	7
e	25	172	13
o	36	11	53

RF LDA max SDS-score	c	e	o
c	478	9	12
e	12	191	7
o	22	6	72

RF LDA max F-measure	c	e	o
c	478	9	12
e	12	191	7
o	22	6	72

DT LDA max SDS-score	c	e	o
c	482	9	8
e	17	187	6
o	21	7	72

DT LDA max F-measure	c	e	o
c	482	9	8
e	17	187	6
o	21	7	72

SVM 15 features max SDS-score	c	e	o
c	490	6	3
e	9	195	6
o	25	5	70

SVM 15 features F-measure	c	e	o
c	490	6	3
e	9	195	6
o	25	5	70

DT 15 features max SDS-score	c	e	o
c	490	5	4
e	8	197	5
o	26	10	64

DT 15 features max F-measure	c	e	o
c	490	5	4
e	8	197	5
o	26	10	64

RF 15 features SDS-score	c	e	o
c	488	7	4
e	7	194	9
o	22	4	74

RF 15 features F-measure	c	e	o
c	489	7	3
e	8	194	8
o	22	3	75

ET 15 features SDS-score	c	e	o
c	487	8	4
e	7	198	5
o	27	5	68

ET 15 features F-measure	c	e	o
c	487	8	4
e	7	198	5
o	27	5	68

GB 15 features SDS-score	c	e	o
c	488	6	5
e	8	194	8
o	20	5	75

GB 15 features F-measure	c	e	o
c	488	6	5
e	8	194	8
o	20	5	75

kNN 15 features SDS-score	c	e	o
c	489	8	2
e	12	193	5
o	30	7	63

kNN 15 features F-measure	c	e	o
c	489	8	2
e	12	193	5
o	30	7	63

MLP 15 features SDS-score	c	e	o
c	485	6	8
e	9	194	7
o	19	5	76

MLP 15 features F-measure	c	e	o
c	485	6	8
e	9	194	7
o	19	5	76

Table B.13: Confusion matrices for the second experiment (raw data Table 7).

Appendix C. Running times of classifiers per 10-fold classification

The average running time of extracting one feature of the cell were 2 ms with standard deviation of 0.8 ms.

Classifier	Features	Average time [s]	Std time [s]
SVM	all	3.17	1.19
	15	0.6	0.15
DT	all	0.35	0.56
	15	0.07	0.04
	PCA	0.12	0.09
	LDA	0.08	0.01
RF	all	4.61	2.45
	15	2.11	1.07
	PCA	1.76	0.46
	LDA	1.46	0.69
ET	all	2.17	1.46
	15	0.89	0.48
GB	all	118.05	58.29
	15	17	7.11
kNN	all	6.69	4.35
	15	0.63	0.55
MLP	all	36.8	18.73
	15	18.79	13.96

Table C.14: Average running times of classifiers (Experiment 1 and Experiment 2), for one iteration grid search cross-validation.

Classifier	Average time [ms]	Std time [ms]
SVM	30	2
DT	3	0.3
RF	30	2
ET	50	6
GB	1068	20
kNN	2	0.2
MLP	2391	45

Table C.15: Training average running times of classifiers, with 15 features only fitting the classifier with training data without cross validation (Experiment 2).