Image Anal Stereol 2024;43:109-119
Original Research Paper

doi: 10.5566/ias.2962

CLASSIFICATION OF RED BLOOD CELLS FROM A GEOMETRIC MORPHOMETRIC STUDY

LLUÏSA GUAL-VAYÀ

Department of Mathematics-IMAC, Higher School of Technology and Experimental Sciences, Universitat Jaume I, 12071-Castelló, Spain
e-mail: al395358@uji.es

(Received July 3, 2023; revised February 13, 2024; accepted February 14, 2024)

ABSTRACT

Sickle cell disease causes the deformation of erythrocytes into sickle cells. The study of this process using digital images of peripheral blood smears can help specialists to quantify the number of deformed cells in order to gauge the severity of the illness. A new method for classifying red blood cells into three categories: healthy, sickle cell disease, and other deformations is proposed. This method does not require obtaining the contour of each cell but instead utilizes information obtained from a small number of points, obtained through appropriate geometric sampling and the use of stereological formulas. The parameters utilized for classification are the bending energy times length (E) and the circular shape factor (F). In normal cells, which exhibit an almost circular shape, these parameters typically have values close to $(1,1)$. To assess the effectiveness of classification using the parameters (E,F), a synthetic curve dataset and a dataset of red blood cells are employed, applying various supervised and unsupervised classification methods.

Keywords: bending energy, cell classification, geometric sampling, integral geometry, stereology.

INTRODUCTION

Normally, red blood cells are flexible and round, moving easily through even the smallest of blood vessels. When a person has sickle cell anemia, many red blood cells assume a rigid sickle-like shape or crescent moon, that can hinder their passage through diminutive capillaries, resulting in oxygen deficiency to certain tissues as blockages form. Moreover, sickled red blood cells are unusually fragile and prone to breakage, so they only survive in the bloodstream for about a tenth of the time that normal erythrocytes remain in circulation, increasing the effects of anemia. Among the most common symptoms of sickle cell anemia are fatigue, breathlessness, joint pain, delayed growth, jaundice, rapid heart rate, increased susceptibility to infections, and sporadic attacks of pain (often termed crises) in the abdomen or other areas of the body.

One method to assess the clinical status of patients is the classification of cells based on their morphology. Cells are generally classified into three categories: normal, sickle, or other abnormalities. Although even today this classification is carried out mostly by specialists, looking directly into the microscope or at the computer screen to decide each cell of what type it is, nowadays there are more and more studies that use automatic cell classification methods, based on image processing techniques and machine learning. Initially, the images are obtained from a peripheral blood smear that gives rise to the microscopic images that are observed by the specialist. However, in order to apply automatic classification techniques, these images are processed and segmented to distinguish the outline of the cells and, in most methods, the outline (flat curve) is used to obtain the classification Sadafi *et al.* (2023). This classification can be made from characteristics that are extracted from the contour (length, area, eccentricity...) (Bischin *el al.* (2012)), or considering the contour (curve) as an element of a more complicated geometric manifold (Epifanio *et al.* (2020)) in which the distances used in the classification are defined or using, for instance, neural nets. The state of the art of segmentation methods and cell classification methods, based on boundary features and geodesic distances in the shape space of curves, can be found at Delgado-Font *et al.* (2020), and a literature review of image processing methods and machine learning methods can be found in Alzubaidi *el al.* (2022). In any case, as we will also see in this work, the classification depends on the segmentation of the cell contours, and fluctuations in the segmentation can strongly influence the classification results.

In this paper we propose a semiautomatic method for classifying red blood cells into three groups: normal, elongated (sickle-shaped) and with other deformations; based on stereological estimators of contour features. The specialists must select certain points of the microscope image of the peripheral blood smear, and from these points the characteristics used in the classification will be estimated. These characteristics are the circular shape factor F (based on the isoperimetric inequality and already used in other works) and the bending energy times length E, which is a new proposal based on the bending energy of the flat curve (contour). The objective of the paper is not to compare our method with other existing methods for red blood cell classification based on cell boundary segmentation. Instead, the aim is to provide a new classification method without the need to segment each cell, but by observing what happens at the intersection of lines with the cell boundary through appropriate geometric sampling. The computational burden of our method depends on the density of lines considered in the geometric sampling (see Fig. 1).

Although stereology is defined as a discipline that allows obtaining efficient and unbiased estimates of 3D quantitative characteristics from an appropriate geometric sampling based on 2D measurements of the structure of interest (Baddeley and Jensen (2005); Howard and Reed (2005)), estimates of 2-dimensional parameters on the plane (area, curve length...) based on sampling with measurements of dimension less than two (lines and points) have also been considered in the literature (see, for instance, Eq. (7.5) and Eq. (7.9) of Baddeley and Jensen (2005)). It is in this sense that we will define stereological estimates of the characteristics F and E.

The paper can be divided into two parts. In the first part (Materials and Methods), the circular shape factor F and the bending energy times length E of a planar closed curve are defined, and two approximations $(\tilde{F},\tilde{E})$ and two stereological estimations $(\hat{F},\hat{E})$ are proposed along with their corresponding square coefficients of error. In the second part, two databases are considered. One consists of thirty synthetic curves for which we know their parametrizations, allowing us to obtain (F,E) exactly. The other is composed of segmented images of peripheral blood smear samples, therefore, we can only obtain approximations and estimations of (F,E). Based on these curve databases, we propose an unsupervised classification of the curves to observe, firstly, if (F,E) provide an adequate classification, and secondly, if this classification remains suitable and is not significantly altered when considering the approximations $(\tilde{F},\tilde{E})$ or estimations $(\hat{F},\hat{E})$.

MATERIALS AND METHODS

Let $\alpha : [a,b] \longrightarrow \mathbb{R}^2$, with $\alpha(t) = \{x(t),y(t)\}$, be a natural regular curve. Let C denote the graph of the curve. We suppose that C bounds a domain D in the plane. Let A and B denote the area of D and the length of C, respectively. The curvature of α at the point $\alpha(t)$ is given by (Gual-Arnau *et al.* (2017))

$$\kappa(t) = \frac{x'y'' - x''y'}{(x'^2 + y'^2)^{\frac{3}{2}}}. \tag{1}$$

The bending energy of the curve α is defined as (Canham (1970), Young *et al.* (1974))

$$E(C) = \int_a^b \kappa^2(t) \|\alpha'(t)\| \, dt. \tag{2}$$

The parameters A, B and $E(C)$ do not depend on the parameterization α, they only depend on the geometrical curve C. The real interval $[a,b]$ defining the curve will be $[0,B]$ when the curve is parameterized by arc length and $[0,2\pi]$ for the closed curves considered in the experimental study.

The bending energy of a closed object in 2D has been frequently used as shape discriminator; in fact, two of the pioneering papers in relating the shape of red blood cells with the bending energy of their boundaries are Canham (1970) and Young *et al.* (1974).

The two features that we will consider in this paper as shape descriptors are the following.

Definition 1. *The circular shape factor of D is defined as*

$$F = \frac{B^2}{4\pi A}. \tag{3}$$

The bending energy times length of C is defined as

$$E = \frac{B\,E(C)}{4\pi^2}. \tag{4}$$

In this paper, the shape of a plane domain will be the geometric information that remains invariant when rotations, translations and/or changes of scale act on the domain. Since the area of a domain and the length and curvature of a curve are invariant under translations and rotations, we have that the circular shape factor F and bending energy times length E will be also invariant. Now we will see that F and E are also invariant under changes of scale.

Let $D_\lambda = \{\lambda(x,y) / (x,y) \in D\}$ with $\lambda > 0$, and C_λ be the curve that bounds D_λ. Then $A(D_\lambda) = \lambda^2 A(D)$, $B(C_\lambda) = \lambda B(C)$ and, from Eq. (1), $\kappa(C_\lambda)(t) = \frac{1}{\lambda}\kappa(C)(t)$; therefore, $E(C_\lambda) = (1/\lambda)E(C)$. We conclude from Eq. (3) and Eq. (4) that E and F are rescaling invariants and therefore shape descriptors.

In fact, from the well-known isoperimetric inequality in the plane, we have that $F \geq 1$ and equality is given if and only if D is a circle.

Proposition 2. *Given a domain D bounded by a differentiable simple closed curve C, then $E \geq 1$ and equality holds if and only if D is a circle.*

Proof of Proposition 2. In the proof, we will use the Fourier series technique employed in Young *et al.* (1974), adapted to our definition of E.

Let C be a differentiable simple closed curve which bounds a planar domain D. Let $\alpha(s) = \{x(s), y(s)\}$ be a parametrization of C by arc length; then $||\alpha'(s)||^2 = x'(s)^2 + y'(s)^2 = 1$, $\forall s \in [0,B]$.

$x(s)$ and $y(s)$ are periodic functions; that is, $x(s+B) = x(s)$ and $y(s+B) = y(s)$; then, we consider the Fourier series of both functions

$$\begin{aligned} x(s) &= \frac{a_0}{2} + \sum_{n=1}^{\infty} \left(a_n \cos\left(\frac{2\pi n}{B}s\right) + b_n \sin\left(\frac{2\pi n}{B}s\right)\right), \\ y(s) &= \frac{c_0}{2} + \sum_{n=1}^{\infty} \left(c_n \cos\left(\frac{2\pi n}{B}s\right) + d_n \sin\left(\frac{2\pi n}{B}s\right)\right). \end{aligned} \tag{5}$$

Since $x'(s)^2 + y'(s)^2 = 1$, $\int_0^B (x'(s)^2 + y'(s)^2) ds = B$; then, from the derivatives of (5), we have

$$B^2 = \sum_{n=1}^{\infty} 2\pi^2 n^2 \left(a_n^2 + b_n^2 + c_n^2 + d_n^2\right). \tag{6}$$

Let $\vec{T}$ be the unit tangent vector to $\alpha(s)$; then, $|\vec{T}'(s)| = \kappa(s)$ and $|\vec{T}'(s)|^2 = x''(s)^2 + y''(s)^2$; then, applying the Parseval's identity to the Fourier series of x', y', x'' and y'' we obtain

$$\begin{aligned} E &= \frac{BE(C)}{4\pi^2} = \frac{B}{4\pi^2}\int_0^B \kappa^2(s) ds \\ &= \frac{B^2}{8\pi^2}\sum_{n=1}^{\infty} \frac{16\pi^4 n^4}{B^4}\left(a_n^2 + b_n^2 + c_n^2 + d_n^2\right). \end{aligned} \tag{7}$$

Having in mind the above equation; to determine the set of Fourier coefficients that yield a minimum E subject to the equality constraint given in Eq. (6) for B^2, we form the function

$$f = E + \lambda\left(B^2 - \sum_{n=1}^{\infty} 2\pi^2 n^2 \left(a_n^2 + b_n^2 + c_n^2 + d_n^2\right)\right), \tag{8}$$

and we apply the method of Lagrange multipliers where the variables will be $z_n^2 = a_n^2 + b_n^2 + c_n^2 + d_n^2$ and the Lagrange multiplier λ; that is

$$f = E + \lambda\left(B^2 - \sum_{n=1}^{\infty} 2\pi^2 n^2 z_n^2\right). \tag{9}$$

Then, from (7),

$$\frac{\partial f}{\partial z_k} = \frac{4\pi^2 k^4}{B^2} z_k - 4\lambda\pi^2 k^2 z_k = 4\pi^2 k^2 z_k \left(\frac{k^2}{B^2} - \lambda\right), \tag{10}$$

and for $\frac{\partial f}{\partial z_k} = 0$ to hold, it must be satisfied that $\lambda = \frac{k^2}{B^2}$, which implies that at most one z_k is nonzero.

From Eq. (6) and Eq. (7), it leads to

$$\frac{E_{min}}{B^2} = \frac{k^2}{B^2}. \tag{11}$$

So, the minimum value of E is given when $k = 1$; that is, $E_{min} = 1$. Moreover, in this case, from Eq. (5) we have

$$\begin{aligned} x(s) &= \frac{a_0}{2} + a_1 \cos\left(\frac{2\pi}{B}s\right) + b_1 \sin\left(\frac{2\pi}{B}s\right), \\ y(s) &= \frac{c_0}{2} + c_1 \cos\left(\frac{2\pi}{B}s\right) + d_1 \sin\left(\frac{2\pi}{B}s\right). \end{aligned} \tag{12}$$

Then, we can write

$$\begin{aligned} &\left(x(s) - \frac{a_0}{2}\right)^2 + \left(y(s) - \frac{c_0}{2}\right)^2 = \\ &= a_1^2 + b_1^2 + c_1^2 + d_1^2 \\ &- a_1^2 \sin^2\left(\frac{2\pi}{B}s\right) - b_1^2 \cos^2\left(\frac{2\pi}{B}s\right) \\ &+ 2a_1 b_1 \cos\left(\frac{2\pi}{B}s\right) \sin\left(\frac{2\pi}{B}s\right) \\ &- c_1^2 \sin^2\left(\frac{2\pi}{B}s\right) - d_1^2 \cos^2\left(\frac{2\pi}{B}s\right) \\ &+ 2c_1 d_1 \cos\left(\frac{2\pi}{B}s\right) \sin\left(\frac{2\pi}{B}s\right). \end{aligned} \tag{13}$$

Differentiating the functions x and y in Eq. (12) and substituting them into $x'(s)^2 + y'(s)^2 = 1$, we can write Eq. (13) as:

$$\begin{aligned} &\left(x(s) - \frac{a_0}{2}\right)^2 + \left(y(s) - \frac{c_0}{2}\right)^2 \\ &= a_1^2 + b_1^2 + c_1^2 + d_1^2 - \frac{B^2}{4\pi^2}. \end{aligned} \tag{14}$$

Now, from Eq. (6),

$$\left(x(s)-\frac{a_0}{2}\right)^2+\left(y(s)-\frac{c_0}{2}\right)^2=\frac{B^2}{2\pi^2}-\frac{B^2}{4\pi^2}=\frac{B^2}{4\pi^2}, \tag{15}$$

which is the equation of a circle of radius $\frac{B}{2\pi}$ centered at $(\frac{a_0}{2},\frac{c_0}{2})$. □

Therefore, to each plane domain D we will associate the pair of points, both greater than or equal to one, (F,E) and whose value will be (1,1) if and only if the domain is a circle.

MATERIALS

The objective of this paper is to characterize a plane shape through the pair (F,E). First we will consider a synthetic base formed by 30 curves of which we know their parameterization α_i (see Fig. 2); therefore, we will be able to obtain the pair (F,E) for each shape exactly. Next we will work with a discretization of the curves, so that we will have a big number of points for each curve and, from them, we will obtain the approximate values $(\tilde{F},\tilde{E})$ of (F,E). Finally, we will use stereological estimators to obtain an estimate $(\hat{F},\hat{E})$ of the values (F,E) of each curve.

Next, we will consider images of peripheral blood smear samples from patients with sickle cell disease in the Special Department of Hematology of the General Hospital 'Dr. Juan Bruno Zayas Alfonso' from Santiago de Cuba. A specialist prepared the blood samples and manually segmented and classified each cell as normal, sickle (elongated) or with other deformations. The dataset used in this study is available at http://erythrocytesidb.uib.es/. From these images we will obtain the approximate values $(\tilde{F},\tilde{E})$ and the stereological estimators $(\hat{F},\hat{E})$ of each red blood cell. We will see that the values of $\tilde{E}$ and especially $\tilde{F}$ depend heavily on the segmentation performed, as variations in the contour strongly affect the curvature and length of the curve. However, the value of the estimations $(\hat{F},\hat{E})$ does not depend as much on these variations in the segmentation, and the values are more stable and similar to those obtained in the synthetic dataset.

METHODS

If we know the parameterizations $\alpha_i : [0,2\pi] \longrightarrow \mathbb{R}^2$ with $\alpha_i(t) = (x_i(t),y_i(t))$, we compute for each curve

$$B=\int_0^{2\pi}\sqrt{(x'(t))^2+(y'(t))^2dt}, \quad A=\int_0^{2\pi}x'(t)y(t)dt, \tag{16}$$

and $\kappa(t)$ from Eq. (1). Then, we associate to each synthetic shape a point in $\mathbb{R}^2$ given by the exact values

$$\left(F=\frac{B^2}{4\pi A},\ E=\frac{BE(C)}{4\pi^2}\right). \tag{17}$$

If we know $\alpha(t)$ in a discrete number of points $\alpha(t_i) = (x(t_i),y(t_i))) = (x_i,y_i)$, $i = 0,1,\dots,N$, where $t_i = \frac{2\pi i}{N}$, the approximate values of A, B and $E(C)$ are obtained from numerical methods. Then we obtain the approximations $(\tilde{F},\tilde{E})$ of (F,E).

To approximate A there exist some numerical methods based, for instance, on the Green's theorem or the trapezoidal rule. Here we use the trapezoidal method, then we have two approximations:

$$\begin{aligned} A &\approx \frac{1}{2}\left|\sum_{i=0}^{N-1}(y_i+y_{i+1})(x_{i+1}-x_i)\right|, \\ A &\approx \frac{1}{2}\left|\sum_{i=0}^{N-1}(x_i+x_{i+1})(y_{i+1}-y_i)\right|. \end{aligned} \tag{18}$$

The length B will be approximated as the sum of the lengths of the segments joining consecutive points $\alpha(t_i)$ and $\alpha(t_{i+1})$; that is,

$$B \approx \sum_{i=0}^{N-1}\sqrt{(x_{i+1}-x_i)^2+(y_{i+1}-y_i)^2}. \tag{19}$$

The curvature κ_i at each point $\alpha(t_i)$ can be approximated following different methods (Gual-Arnau *et al.* (2017)), we consider the following,

$$\tilde{\kappa}_i=\frac{4A(T_i)}{a_ib_ic_i}, \tag{20}$$

where T_i is the triangle formed by the points $\alpha(t_{i-1})$, $\alpha(t_i)$ and $\alpha(t_{i+1})$ and a_i,b_i,c_i the lengths of the triangle sides. Now, $E(C)$ is approximated as

$$E(C) \approx \sum_{i=0}^{N-1}\tilde{\kappa}_i^2\sqrt{(x_{i+1}-x_i)^2+(y_{i+1}-y_i)^2}. \tag{21}$$

Finally, to use stereological estimators, we need a suitable geometric sampling. We will consider a square grid of test lines which is isotropic uniform random (IUR) relative to the cells to be estimated. A square grid of test lines is the union of two mutually perpendicular IUR series of parallel test lines a constant distance $T > 0$ apart (see Fig. 1). In practice, isotropic uniform randomness is attempted by superimposing the grid ”at random”, without looking at the image.

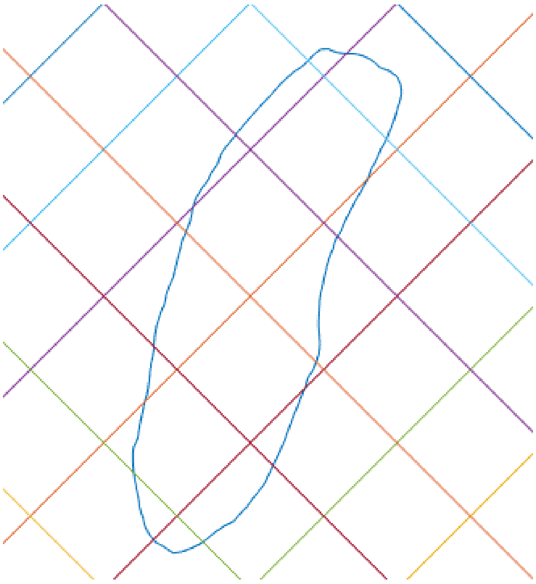


Fig. 1. *A square grid of test lines and a cell in* $\mathbb{R}^2$.

Unbiased estimators of A, B and $E(C)$, in the line, for example, of equations (4.4), (4.5), (4.38), (7.5), and (7.9) in Baddeley and Jensen (2005), are:

$$\begin{aligned} \hat{A} &= T^2 P, \\ \hat{B} &= \frac{\pi}{4} T I, \\ \hat{E}(C) &= \frac{\pi}{4} T \sum_{i=1}^{I} \kappa_i^2, \end{aligned} \tag{22}$$

where P is the number of test points hitting D, I the number of intersection points of lines with $C = \partial D$ and κ_i denotes the curvature of C at an intersection point of C with a line. In Fig. 1, $P = 3$ and $I = 8$. In practice, κ_i is substituted by $\tilde{\kappa}_i$.

Then, the estimates of F and E are, respectively,

$$\left(\hat{F} = \frac{\hat{B}^2}{4\pi \hat{A}}, \, \hat{E} = \frac{\hat{B}\hat{E}(C)}{4\pi^2} \right). \tag{23}$$

The error variance predictors of the above estimators are detailed in Appendix A.

RESULTS

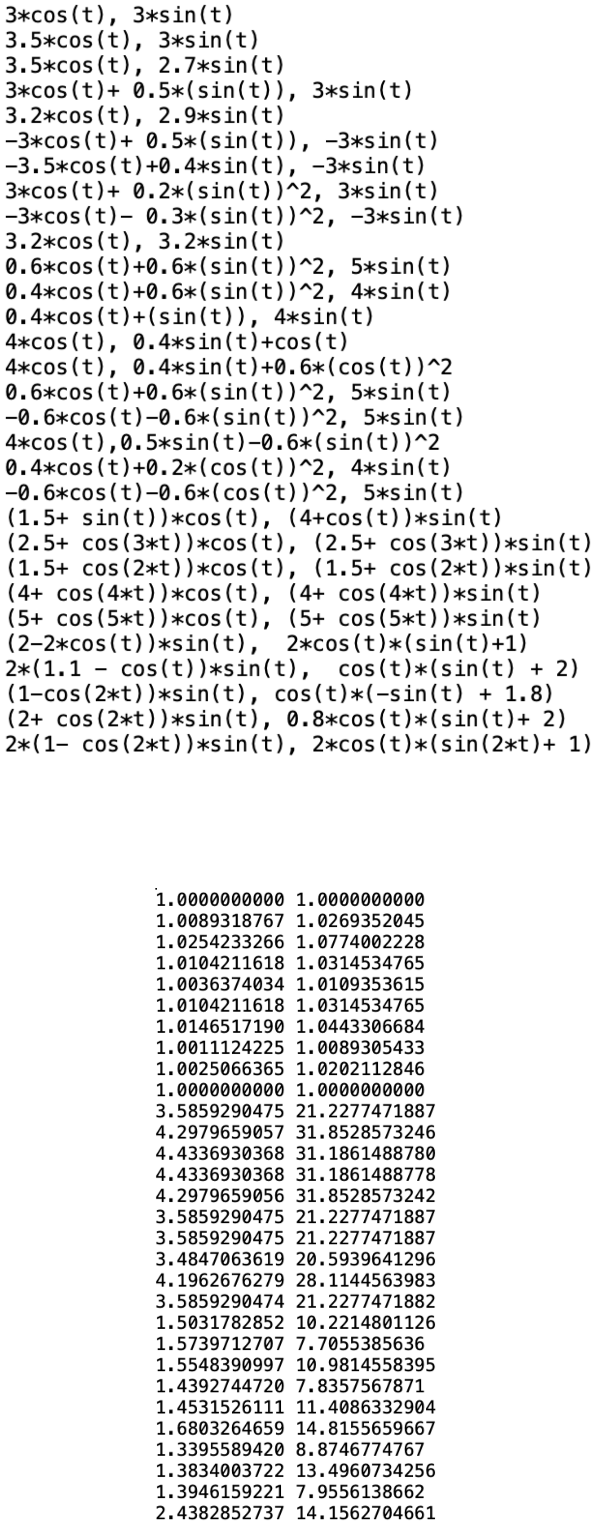
Classification is a type of machine learning task that involves training an algorithm to identify which category or class an observation belongs to. Classification methods are divided into supervised and unsupervised methods. The cluster analysis, also known as unsupervised classification, is a technique used to identify natural groupings or clusters in a dataset based on the values of one or more variables; in our case in the variables F and E. There exist several unsupervised classification methods. We consider a model-based classification method provided by the Expectation-Maximization (EM) algorithm using the MClust library in R, to see if the unsupervised classification we obtain with the factors (F,E), and their approximations and estimations aligns with the known classification. On the other hand, the calculation of $(\tilde{F},\tilde{E})$ and $(\hat{F},\hat{E})$ has been carried out using custom codes developed in MATLAB.

EXPERIMENTAL STUDY OF SYNTHETIC FIGURES

In this section we consider the set of synthetic curves given in Fig. 2 and we apply the classification method from the values of (F,E).

In Fig. 3, we have an example of each class into which we have divided the synthetic figures.

The classification algorithm divides the cells using the factors (F,E) into three distinct groups. In the first group we have the normal (‘almost circular’) cells, whose values of (F,E) are close to (1,1), the sickle cells form the second group and, finally, we have the group of other cells (see Fig. 4).



```
3*cos(t), 3*sin(t)
3.5*cos(t), 3*sin(t)
3.5*cos(t), 2.7*sin(t)
3*cos(t)+ 0.5*(sin(t)), 3*sin(t)
3.2*cos(t), 2.9*sin(t)
-3*cos(t)+ 0.5*(sin(t)), -3*sin(t)
-3.5*cos(t)+0.4*sin(t), -3*sin(t)
3*cos(t)+ 0.2*(sin(t))^2, 3*sin(t)
-3*cos(t)- 0.3*(sin(t))^2, -3*sin(t)
3.2*cos(t), 3.2*sin(t)
0.6*cos(t)+0.6*(sin(t))^2, 5*sin(t)
0.4*cos(t)+0.6*(sin(t))^2, 4*sin(t)
0.4*cos(t)+(sin(t)), 4*sin(t)
4*cos(t), 0.4*sin(t)+cos(t)
4*cos(t), 0.4*sin(t)+0.6*(cos(t))^2
0.6*cos(t)+0.6*(sin(t))^2, 5*sin(t)
-0.6*cos(t)-0.6*(sin(t))^2, 5*sin(t)
4*cos(t),0.5*sin(t)-0.6*(sin(t))^2
0.4*cos(t)+0.2*(cos(t))^2, 4*sin(t)
-0.6*cos(t)-0.6*(cos(t))^2, 5*sin(t)
(1.5+ sin(t))*cos(t), (4+cos(t))*sin(t)
(2.5+ cos(3*t))*cos(t), (2.5+ cos(3*t))*sin(t)
(1.5+ cos(2*t))*cos(t), (1.5+ cos(2*t))*sin(t)
(4+ cos(4*t))*cos(t), (4+ cos(4*t))*sin(t)
(5+ cos(5*t))*cos(t), (5+ cos(5*t))*sin(t)
(2-2*cos(t))*sin(t),  2*cos(t)*(sin(t)+1)
2*(1.1 - cos(t))*sin(t),  cos(t)*(sin(t) + 2)
(1-cos(2*t))*sin(t), cos(t)*(-sin(t) + 1.8)
(2+ cos(2*t))*sin(t), 0.8*cos(t)*(sin(t)+ 2)
2*(1- cos(2*t))*sin(t), 2*cos(t)*(sin(2*t)+ 1)
```

```
1.0000000000 1.0000000000
1.0089318767 1.0269352045
1.0254233266 1.0774002228
1.0104211618 1.0314534765
1.0036374034 1.0109353615
1.0104211618 1.0314534765
1.0146517190 1.0443306684
1.0011124225 1.0089305433
1.0025066365 1.0202112846
1.0000000000 1.0000000000
3.5859290475 21.2277471887
4.2979659057 31.8528573246
4.4336930368 31.1861488780
4.4336930368 31.1861488778
4.2979659056 31.8528573242
3.5859290475 21.2277471887
3.5859290475 21.2277471887
3.4847063619 20.5939641296
4.1962676279 28.1144563983
3.5859290474 21.2277471882
1.5031782852 10.2214801126
1.5739712707 7.7055385636
1.5548390997 10.9814558395
1.4392744720 7.8357567871
1.4531526111 11.4086332904
1.6803264659 14.8155659667
1.3395589420 8.8746774767
1.3834003722 13.4960734256
1.3946159221 7.9556138662
2.4382852737 14.1562704661
```

Fig. 2. *(a) Parametrizations of the synthetic shapes. (b) Exact values of* (F,E).

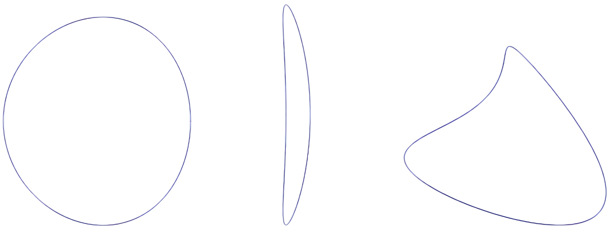


Fig. 3. *Synthetic figures of simulated cells. (a) Normal. (b) Sickle. (c) Other.*

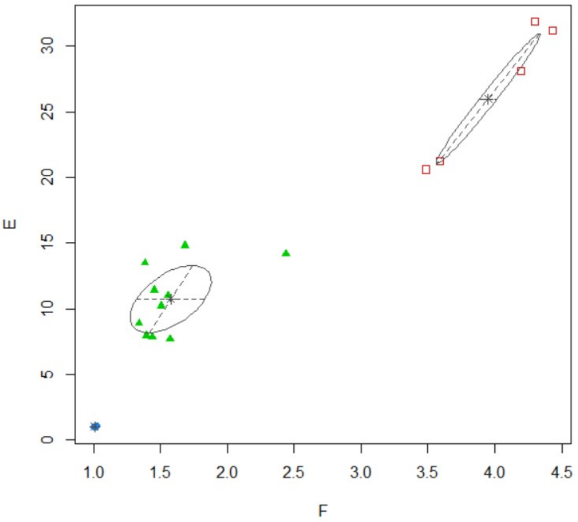


Fig. 4. *MClust method for data classification from* (F,E).

The cluster vector obtained using the MClust library is provided in Eq. (24). This vector indicates the group to which each curve in Fig. 2 has been classified. Therefore, considering the shapes of the synthetic figures, the result aligns with expectations, with ten cells in each group. To provide a rough characterization of the defining features of each group, we have included the average F and E values for each group in Table 1.

$$1111111111\ 2222222222\ 3333333333 \tag{24}$$

Table 1. *Average F and E values of each group.*

	Normal	**Sickle**	**Other**
F	1.01	3.95	1.58
E	1.02	25.97	10.74

Using the classification method provided by the MClust library in R and utilizing the approximate values $(\widetilde{F},\ \widetilde{E})$, we also derive the cluster vector presented in Eq. (24). Consequently, a classification with ten cells in each expected group is achieved (refer to Fig. 5).

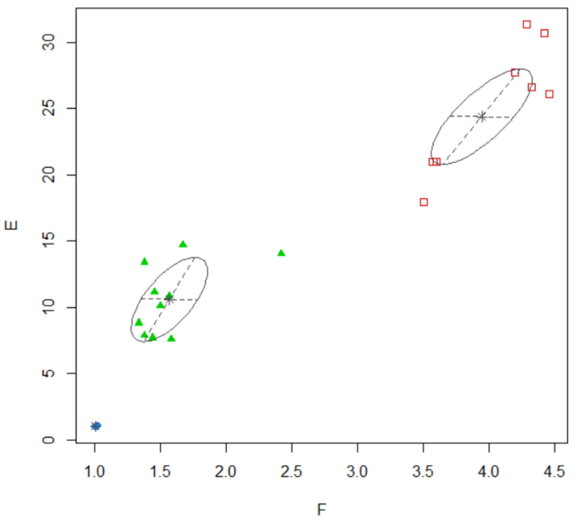


Fig. 5. *MClust method for data classification from* $(\tilde{F}, \tilde{E})$.

In Table 2 we have the average $\tilde{F}$ and $\tilde{E}$ values of each group.

Table 2. *Average* $\tilde{F}$ *and* $\tilde{E}$ *values of each group.*

	Normal	Sickle	Other
F	1.01	3.95	1.57
E	1.01	24.44	10.61

Now, we consider the unbiased estimators of A, B and $E(C)$ derived from Eq. (22). To classify the estimates of F and E, a code draws the parameterizations and the grid of UR lines. The estimation of length and area has been manually calculated by counting the interior points and intersection points; and then, the value of $\hat{F}$ has been obtained. To calculate $\hat{E}$, we must find the sum of squared curvatures at each intersection point of the grid with the curve. To do this, the corresponding points must be clicked, and when all points are clicked, a code calculates $\hat{E}$. To calculate the curvature at a point, we need the point itself and two neighboring points. For this reason, we have only calculated the estimated ($\hat{F}$, $\hat{E}$) values for twelve cells.

As our database is segmented, we have utilized the segmentation points. However, since the curvature depends on the segmentation, our proposal for the future is to not segment the image and have the specialist mark only three points at each intersection between curve and straight line.

We have already seen that the classification of the synthetic figures in Table 2 using the real values of F and E, as well as the approximate values $(\tilde{F}, \tilde{E})$, yields the expected results. However, with stereological estimations, the classification will depend on the errors made in these estimations, specifically with the number of lines used and therefore the value of T. We will only use a fixed value of T, and we will give a predicted approximations of the square coefficients of error of $\hat{F}$ and $\hat{E}$. It is possible to adjust the value of T as we please or depending on how precise we want the estimation to be.

In our case, using the line density shown in Fig. 1, and based on the classification method provided by the MClust library in R, employing the estimated values $(\widehat{F},\ \widehat{E})$, we also obtain the cluster vector presented in Eq. (24). (see Fig. 6).

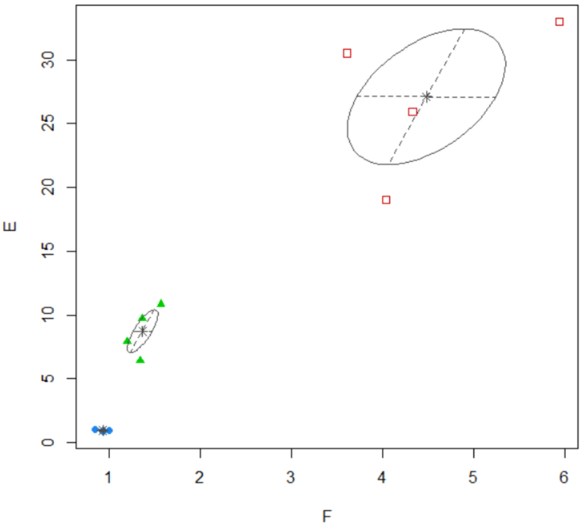


Fig. 6. *MClust method for data classification from* $(\hat{F}, \hat{E})$.

In Table 3 we have the average $\hat{F}$ and $\hat{E}$ values of each group.

Table 3. *Average* $\hat{F}$ *and* $\hat{E}$ *values of each group.*

	Normal	Sickle	Other
F	0.94	4.48	1.37
E	0.96	27.12	8.73

Therefore, given the shapes of the synthetic figures, the classification outcome is as expected, with ten cells in each group, whether we use E and F or their approximations or estimates. Furthermore, the probability that each curve belongs to the group to which it has been classified consistently exceeds

0.9999987; thus, the classification is accurate, well-separated, and correct for all three groups in all cases.

Since we know the exact values of F and E for synthetic figures, we can obtain the sample variance and squared error coefficient of $\hat{F}$ and $\hat{E}$ for each curve, and we have obtained the following results:

For nearly circular curves: $\mathrm{ce}^2(\hat{F}) < 3\%$ and $\mathrm{ce}^2(\hat{E}) < 3\%$.

For sickle curves: $\mathrm{ce}^2(\hat{F}) < 10\%$ and $\mathrm{ce}^2(\hat{E}) < 10\%$.

For the group of other curves: $\mathrm{ce}^2(\hat{F}) < 5\%$ and $\mathrm{ce}^2(\hat{E}) < 5\%$.

SEGMENTED CELLS

In this section we first consider a database from the Department of Hematology of the General Hospital 'Dr. Juan Bruno Zayas Alfonso' from Santiago de Cuba, formed by 513 cells of which 202 are normal, 100 have the sickle cell disease and 211 present other deformations. The initial idea is to use the approximations and estimations of F and E for performing an unsupervised classification of these cells. In Wheeless *et al.* (1994) the circular shape factor is used for the same purpose.

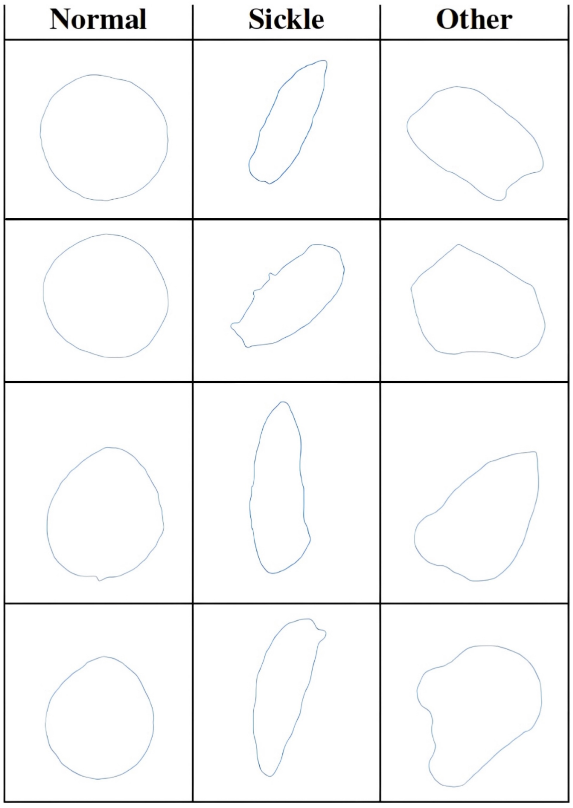


Fig. 7. *Selected figures of segmented red blood cells.*

However, when performing a classification process, it is advisable to remove from the database those elements considered atypical. In this regard, if we observe Fig. 7, we can see that the length of a curve and, especially, the integral of the squared curvature are highly sensitive to curve perturbations; therefore, especially the value of E can vary abruptly with curve perturbations. Therefore, before initiating the classification process with the complete database, we have proceeded to eliminate elements with $F > 92.211$ or $E > 92.229$. As a result, we are left with a database consisting of 202 normal cells and 100 sickle cells, as well as 192 cells with other deformations.

In Fig. 8, we have one erythrocyte from each class of the database and their segmentation (in blue). To obtain the values of $\tilde{E}$ and $\tilde{F}$, complete segmentation is necessary. However, the values of $\hat{E}$ and $\hat{F}$ do not require the entire segmentation but only the values at the intersection points of the boundary with a square grid of test lines.

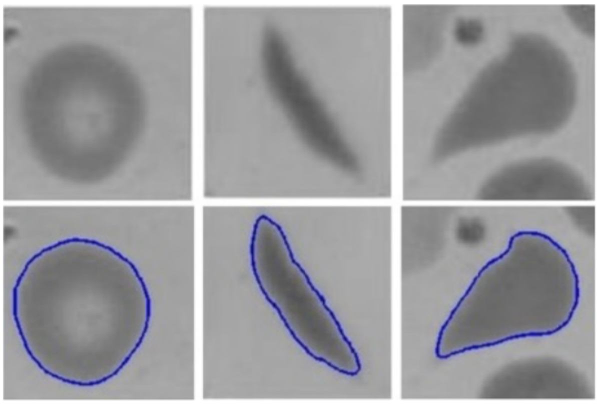


Fig. 8. *Some types of erythrocytes from the database.*

In this section, we are going to calculate the approximate descriptors of E and F for the cells corresponding to the refined and outlier-free database. Let's remember that this is a real database, and therefore we do not have the curves given as parameterizations but as a set of points. The code used to calculate the approximate $\tilde{F}$ and $\tilde{E}$ values is the same as we used in the previous section when calculating a series of points for each parameterization and working with them.

Using Mclust, we have three groups, each composed of 202, 106, and 186 cells, representing normal, sickle, and other cells, respectively. In Fig. 9, we can observe how the groups have been classified. As we can see, three groups emerge. The 202 cells considered normal appear in a single group; therefore, all normal cells are classified within a group where cells from the other two groups do not appear. Then we have a group where the majority cells are sickle cells but some 'other' cells are also classified in this group. The last group is formed by other cells.

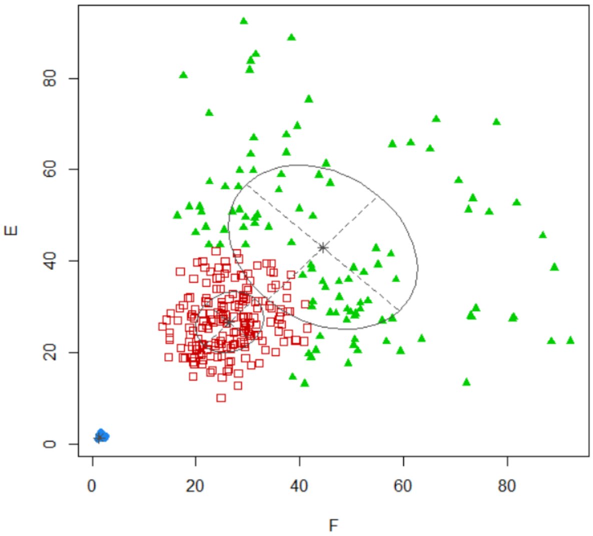


Fig. 9. *MClust method for database classification from approximations.*

Finally, we need to calculate the $\hat{E}$ and $\hat{F}$ values for each cell using stereology. The code to be used is the same as in the preceding section, with the only difference that now we cannot draw the cell in order to click on the intersection points and calculate the two descriptors $\hat{E}$ and $\hat{F}$. The solution to this problem has been to use the scatter function provided by MATLAB, which draws the contour formed by connecting the given points. Therefore, we can use the same code, but using also the code that draw the cell.

We have considered the cell dataset after the removal of atypical elements, meaning we start with the database in which 202 cells are normal, 100 cells are sickle cells, and 192 exhibit other deformations. As a result, we have calculated the estimated $\hat{E}$ and $\hat{F}$ for all the cells in this dataset and subsequently classified them. Applying MClust again, we obtained three groups, one composed of 202 normal cells, a second group consisting of 127 cells classified as sickle cells, and a third group with 165 cells classified with other deformations. Five sickle cells were classified in the group of other deformations, and 32 cells that belonged to the group of other deformations were classified as sickle cells.

The classification results of the MClust method based on the estimates depend on the grid of test lines used and their density. In our case, the grid of lines used and one of the cells can be seen in Fig. 1, and the classification results are shown in Fig. 10.

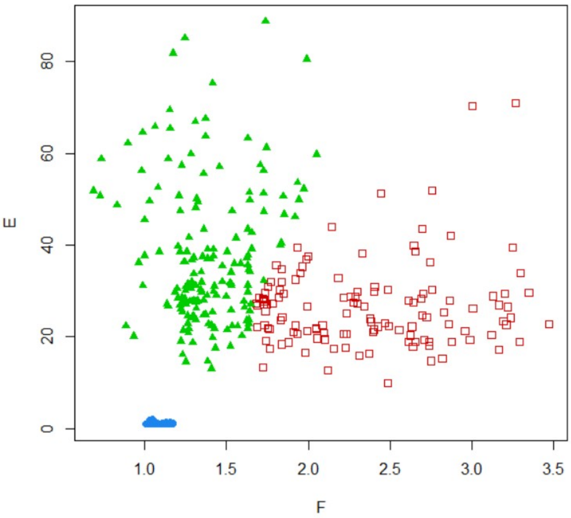


Fig. 10. *MClust method for database classification from estimations.*

In this case, since we do not have the exact values of F and E, we will approximate the squared error coefficient of $\hat{F}$ and $\hat{E}$ of the cells of Fig. 8 using the approximations from the Appendix, and we obtain the following results:

For the normal cell: $\mathrm{ce}^2(\hat{F}) = 2.69\%$ and $\mathrm{ce}^2(\hat{E}) = 1.97\%$.

For the sickle cell: $\mathrm{ce}^2(\hat{F}) = 9.39\%$ and $\mathrm{ce}^2(\hat{E}) = 13.29\%$.

For the other group cell: $\mathrm{ce}^2(\hat{F}) = 9.11\%$ and $\mathrm{ce}^2(\hat{E}) = 8.51\%$.

DISCUSSION

In this paper, we have proposed two parameters, one known as the circular shape factor F and a new one based on the curvature of the curve, the bending energy times length E, which, when associated with red blood cells, have allowed for the classification of these cells into three groups: healthy, with sickle cell disease, and with other deformations. Furthermore, it has been demonstrated that both the values of F and E are greater than one, and when they take the value of 1, they characterize a circle.

The classification has been carried out using both supervised and unsupervised methods, and it has been found that when the values of F and E are replaced with approximations $\tilde{F}$ and $\tilde{E}$ or estimations $\hat{F}$ and $\hat{E}$, the classification does not vary substantially.

The idea would be to incorporate the calculation of $\hat{F}$ and $\hat{E}$ into interactive stereology programs adapted to microscopes so that, by directly observing the microscope images without segmenting the cells, these estimations could be obtained. In this study, since we had a database of segmented images, we used them to compare the results with those obtained using the estimations $\tilde{F}$ and $\tilde{E}$. The two factors F and E vary with fluctuations in the cell segmentation, especially the factor E heavily relies on variations in the contour, as it is based on the curvature of the curve. Therefore, approximating the curvature at a few points directly on the microscope, without the need to segment all the cells, could be better for both the microscope observer and the cell classification.

This way of classifying cells based on information from a small number of points on the boundary can also be useful when there is overlap between cells and it is difficult to obtain the segmentation of each one separately.

APPENDIX

To give an expression of the variance of estimators $\hat{B}$ and $\hat{E}(C)$ in Eq. (22) we are going to give a new expression of both estimators.

Let n_1 be the number of parallel lines in one direction that intersect the curve C and n_2 the number of lines perpendicular to the previous ones that intersect C. Let

$$\{I_{i1}, I_{i2}, \ldots, I_{in_i}\}, \qquad i = 1, 2, \tag{25}$$

denote the total numbers of intersections determined in the curve by the lines hitting the curve. We denote

$$\hat{Q} = \frac{\pi}{4} T \sum_{i=1}^{2} \sum_{j=1}^{n_i} f_{ij}. \tag{26}$$

Then, $\hat{Q} = \hat{B}$ if $f_{ij} = I_{ij}$ (see Eq. (12) of Gómez *et al.* (2016)) and $\hat{Q} = \hat{E}(C)$ if $f_{ij} = \sum_{k=1}^{I_{ij}} \kappa_k^2$.

A predictor of $\operatorname{Var}(\hat{Q})$ from a single IUR superimposition of the grid and the curve is (Eq. (5.7.5) of Cruz-Orive (2024)),

$$\operatorname{var}(\hat{Q}) = \frac{\pi^2 T^2}{96} \left[\left(\sum_{j=1}^{n_1} f_{1j} - \sum_{j=1}^{n_2} f_{2j} \right)^2 + \frac{5}{12} \hat{\nu} \right], \tag{27}$$

where, for $n_i \geq 3$,

$$\hat{\nu} = \sum_{i=1}^{2} (3C_{0i} - 4C_{1i} + C_{2i}), \tag{28}$$

and

$$C_{ki} = \sum_{j=1}^{n_i - k} f_{ij} f_{ij+k}, \quad k = 0, 1, \ldots, n_i - 1; \quad i = 1, 2. \tag{29}$$

On the other hand, when considering the Taylor expansion of the function $(\hat{B})^2$ around the mean value of $\hat{B}$, $\mathbb{E}(\hat{B})$, the following approximations to order $\operatorname{Var}(\hat{B})$ are obtained (for more details, refer to Section 4.3.2 of Benaroya *et al.* (2005)):

$$\begin{aligned} \mathbb{E}(\hat{B})^2 &\approx \mathbb{E}^2(\hat{B}) + \operatorname{Var}(\hat{B}), \\ \operatorname{Var}(\hat{B})^2 &\approx 4\mathbb{E}^2(\hat{B}) \operatorname{Var}(\hat{B}). \end{aligned} \tag{30}$$

Then, we have the following predictors of the square coefficients of error of $(\hat{B})^2$, $\hat{A}$ and $\hat{E}(C)$;

$$\begin{aligned} \operatorname{ce}^2((\hat{B})^2) &= \frac{\operatorname{Var}((\hat{B})^2)}{\mathbb{E}^2((\hat{B})^2)} = \frac{4\mathbb{E}^2(\hat{B}) \operatorname{Var}(\hat{B})}{(\mathbb{E}^2(\hat{B}) + \operatorname{Var}(\hat{B}))^2}. \\ \operatorname{ce}^2(\hat{A}) &= \frac{\operatorname{Var}(\hat{A})}{\mathbb{E}^2(\hat{A})} = 0.07284 \frac{\hat{B}}{\sqrt{\hat{A}} P^{\frac{3}{2}}}. \\ \operatorname{ce}^2(\hat{E}(C)) &= \frac{\operatorname{Var}(\hat{E}(C))}{\mathbb{E}^2(\hat{E}(C))}. \end{aligned} \tag{31}$$

The second equation can be found in Gundersen and Jensen (1987). Finally, using Cochran's formula and Goodman's formula, respectively, and supposing independent estimators of A, B, and $E(C)$, we obtain the following predicted approximations of the squared coefficients of error for $\hat{F}$ and $\hat{E}$ (Cruz-Orive (2024), Sections A.2.3 and A.2.4):

$$\begin{aligned} \operatorname{ce}^2(\hat{F}) &= \operatorname{ce}^2((\hat{B})^2) + \operatorname{ce}^2(\hat{A}). \\ \operatorname{ce}^2(\hat{E}) &= \operatorname{ce}^2(\hat{B}) + \operatorname{ce}^2(\hat{E}(C)). \end{aligned} \tag{32}$$

ACKNOWLEDGMENDS

The author would like to express gratitude to the Associate Editor and the four anonymous reviewers for their highly constructive suggestions, which have significantly enhanced the manuscript. This research has been funded through a research grant from the Institute of Mathematics and Applications (IMAC) at Jaume I University (UJI).

REFERENCES

Alzubaidi L, Fadhel MA Al-Shamma O, Zhang J, Duan Y (2022). Deep Learning Models for Classification of Red Blood Cells in Microscopy Images to Aid in Sickle Cell Anemia Diagnosis. Electronics 9:1–18.

Baddeley A, Vedel-Jensen EB (2005). Stereology for Statisticians, 1st ed. Chapman & Hall.

Benaroya H, Han SM, Nagurka M (2005). Probability Models in Engineering and Science, CRC Press, Taylor and Francis Group.

Bischin C, Ţălu Ş, Silaghi-Dumitrescu R, Ţălu M, Giovanzana S, Lupaşcu CA (2012). Computerized morphometric assessment of the human red blood cells treated with cisplatin. Ann Rom Soc Cell Biol 17(2): 105-10.

Canham PB (1970). The Minimum Energy of Bending as a Possible Explanation of the Biconcave Shape of the Human Red Blood Cell. J Theor Biol 26:61–81.

Cruz-Orive LM (2024). Stereology. Theory and Applications, IAM Series, Springer.

Delgado-Font W, Escobedo-Nicot M, González-Hidalgo M, Herold-García S, Jaume-i-Capó A, Mir A (2020). Diagnosis support of sickle cell anemia by classifying red blood cell shape in peripheral blood images. Med Biol Eng Comput 58:1265–84.

Epifanio I, Gual-Arnau X, Herold S (2020). Morphological analysis of cells by means of an elastic metric in the shape space. Image Anal Stereol 39(1):13–23.

Gómez AI, Cruz M, Cruz-Orive LM (2016). On the precision of curve length estimation in the plane. Image Anal Stereol 35(1):1–14.

Gual-Arnau X, Ibáñez Gual MV, Monterde J (2017). Curvature approximation from parabolic sectors. Image Anal Stereol 36(3):233–41.

Gundersen HJG, Jensen EB (1987). The efficiency of systematic sampling in stereology and its prediction. J Microsc-Oxford 147:229–63.

Howard V, Reed MG (2005) Unbiased Stereology, 2nd ed.; Garland science/BIOS Scientific Publishers, Oxford: England.

Sadafi A, Bordukova M, Makhro A, Navab N, Bogdanova A, Marr C (2023). RedTell: an AI tool for interpretable analysis of red blood cell morphology. Front Physiol 14:1058720.

Wheeless LL, Robinson RD, Lapets OP, Cox C, Rubio A, Weintraub, M, Benjamin, LJ (1994). Classification of Red Blood Cells as Normal, Sickle, or Other Abnormal, Using a Single Image Analysis Feature. Cytometry 17:159-66.

Young IT, Walker JE, Bowie JE (1974). An Analysis Technique for Biological Shape. Inform Control 25:357–70.