Mathematical and Computational Applications

Article

Classification of Red Blood Cells in the Kendall Space of Reflection Shapes

Ximo Gual-Arnau [1,*,†] and Lluïsa Gual-Vayà [2,†]

[1] Departament de Matemàtiques, Institute of New Imaging Technologies, Universitat Jaume I, 12071 Castelló, Spain

[2] Departament de Matemàtiques, Institute of Mathematics and Applications, Universitat Jaume I, 12071 Castelló, Spain; al395358@uji.es

[*] Correspondence: gual@uji.es

[†] The authors contributed equally to this work.

Abstract: The classification of red blood cells (RBCs) or erythrocytes into three categories based on their shape, normal, sickle-shaped, and those with other deformations, has proven to be a crucial tool in diagnosing and managing sickle cell disease (SCD). Manual classification techniques have evolved into automated tools, with numerous classification methods being applied based on different ways of representing the cells. In this work, we propose a novel methodology for representing RBCs, defined by selecting k landmarks along the cell boundaries and characterizing shapes as points in the Kendall space of reflection shapes, Ω_2^k. Using this representation, we applied an embedding of the Kendall space Ω_2^k into a Euclidean space, which allowed for the use of machine learning classification algorithms. We also compared our results with those obtained using other classification methods applied to the same dataset in the literature, highlighting the strong performance of our approach in terms of classification accuracy.

Keywords: erythrocytes; Kendall space; machine learning algorithms; shape classification

Citation: Gual-Arnau, X.; Gual-Vayà, L. Classification of Red Blood Cells in the Kendall Space of Reflection Shapes. *Math. Comput. Appl.* **2024**, *29*, 122. https://doi.org/10.3390/mca29060122

Academic Editor: Leonardo Trujillo

Received: 1 October 2024
Revised: 12 December 2024
Accepted: 17 December 2024
Published: 19 December 2024

1. Introduction

Sickle cell disease (SCD) is a severe genetic blood disorder characterized by the presence of an abnormal form of hemoglobin, hemoglobin S (HbS), which causes red blood cells (RBCs) to adopt a rigid crescent or sickle shape. RBCs, which are normally circular and flexible, play a vital role in transporting oxygen throughout the body by moving easily through tiny blood vessels. However, in individuals with SCD, the deformed sickle-shaped cells obstruct normal blood flow, leading to pain and other complications in various parts of the body. While there are many blood abnormalities, SCD stands out due to its significant impact on millions of people worldwide. Diagnosing SCD relies on the classification of RBCs based solely on their shape, distinguishing between normal cells, sickle cells, and other deformed shapes, without consideration of the cell size. This shape-based classification is essential for identifying the presence of sickle cells, assessing disease severity, monitoring progression, and guiding timely interventions to improve the quality of life for affected individuals.

The automatic classification of red blood cells based on their shape in images of peripheral blood smear samples is a field that has developed significantly in recent years through various mathematical and computational approaches. One of the approaches involves using simple descriptors such as the circularity, ellipticity, or bending energy [1,2] to describe the shape of the cells, and then applying classification methods based on these descriptors. The utilization of descriptors based on Fourier series or the template matching technique has also been proposed [3]. Artificial neural networks have been employed [4], as well as shape features based on integral geometry [5]. In [6], a review of segmentation and classification methods for red blood cells can be found. A recent model for classifying red blood cells that we want to highlight involves considering the contour of each cell as a parametrized curve in $\mathbb{R}^2$. In this way, metrics in the shape space, which is formed by all closed parametrized curves and has the structure of a Riemannian manifold, have been used for classification [7,8]. However, in this model, where each shape is represented by a parametrized curve, achieving the invariance of the curve (shape) to motions and scaling is more feasible, while obtaining shape invariance against changes in parametrization is a much more mathematically and computationally expensive process. Therefore, considering that when working with databases, the boundary curves of erythrocytes are discrete (formed by a finite number of points), in this work, we introduce the representation of shapes using a set of landmarks from the boundary curve. This approach achieves excellent classification results while reducing the computational cost.

At present, both machine learning and deep learning are two mainstream approaches for classification. Regarding deep learning, transfer learning models such as lightweight models, ResNet-50, AlexNet, and VGG have been implemented for RBC classification [9,10]. In this work, as we will explain below, we present a novel representation of the red blood cell shape that allows us to apply machine learning classification algorithms. We also compare our classification results with those obtained using other methods, including deep learning approaches, applied to the same dataset.

All the previous studies highlight the growing use of advanced machine learning algorithms in cell classification tasks, demonstrating promising results in automated analysis and diagnostic support. Our work introduces a theory that, although well known, is entirely novel in the classification of red blood cells. This theory enables the introduction of new mathematical tools and SVM classification methods tailored to the specific challenges of red blood cell morphology.

In numerous applications, planar shapes are characterized by a finite number of points along their contours, and the required geometric information in a shape must also be invariant under reflections of the landmark set. In this work, we represented each planar shape (cell) using a set of digitized points that described its contour, provided directly by the dataset. These points were evenly distributed along the cell boundary, ensuring a detailed and uniform representation of each shape. To define the landmarks required for working in Kendall's reflection shape space [11], we adopted a mathematical approach based on the geometry of the cell. Specifically, the mathematical landmarks were defined as the two boundary points that determined the cell's diameter (the pair of points with the maximum Euclidean distance between them). The remaining points, referred to as pseudo-landmarks, were distributed consistently along the boundary relative to the identified diameter. This approach ensured a systematic correspondence between points across different shapes, even in the absence of anatomical homology. In our application, we aimed to classify red blood cells into three categories: normal, sickle-shaped, and those with other deformations. Each cell was thus described by k landmarks in $\mathbb{R}^2$, which corresponded consistently across all cells.

The reason we considered invariance to reflections and, consequently, worked in the Kendall space of reflection shapes is that the blood images used were taken from samples prepared by smearing or spreading. As a result, the same red blood cell in the three-dimensional blood sample, if rotated 180 degrees around an axis, can produce a two-dimensional projection that is a reflection shape of the one obtained without rotating the red blood cell. Therefore, the 2D reflections do not alter the shape of the cell.

We analyzed a dataset comprising 202 normal cells, 210 sickle cells, and 211 cells with other cellular deformations. These cells with other deformations, although they also exhibit morphological alterations, are not directly relevant to sickle cell disease and their specific identification does not provide additional information for the diagnosis, monitoring, or treatment of the disease. Therefore, they were grouped into a generic category of "other deformations". In line with the results of [12], but considering shape invariance with respect to the orthogonal group $O(2)$ instead of the special orthogonal group $SO(2)$, we considered the embedding of the Kendall space of reflection shapes Ω_2^k into the Euclidean space $\mathbb{R}^{k(k-1)/2}$. Then, we considered the extrinsic distance induced by the Euclidean distance of this embedding and the Euclidean distance transformed by the kernel, resulting from the embedding of Ω_k^2 into a reproducing kernel Hilbert space. Using these distances, we applied various machine learning algorithms to perform the supervised classification and unsupervised clustering of the dataset and conduct a classification evaluation.

On the other hand, we present a new identification of the submanifold $M = \mathrm{i}(\Omega_k^2)$ within $\mathbb{R}^{k(k-1)/2}$. Additionally, we defined a new sample mean on M, derived from the extrinsic sample mean in $\mathbb{R}^{k(k-1)/2}$ using the Euclidean distance. The corresponding shape was then obtained via the inverse function i^{-1}.

This paper is organized as follows. In Section 2, we present an overview of the Kendall space of reflection shapes Ω_2^k and introduce the new methodology we used for the classification of red blood cells. This methodology involved an embedding and a novel characterization of the Ω_2^k space. We also explain how each cell in the database was represented as a point in Ω_2^k and detail the classification algorithms applied in both supervised classification and unsupervised clustering. Finally, in Sections 3 and 4, we discuss the classification results obtained and present the conclusions of the study, respectively.

2. Materials and Methods

In this section, we begin by reviewing Kendall's reflection shape space Ω_2^k [11]. Next, in Theorem 1, we define an embedding of Ω_2^k into a Euclidean space, inspired by [12], but also considering invariance under reflections. The main novelty is in Proposition 1, which provides a new characterization of the space Ω_2^k, from which we define a new extrinsic mean. Finally, we detail the cell dataset used in this work and the proposed classification methods.

2.1. Kendall Spaces

Although the erythrocyte size is important in various blood pathologies, in the case of sickle cell disease, the detection of sickle cells primarily depends on their shape. Therefore, the classification of red blood cells into the three categories was based solely on the shape of the cells, without considering their size.

Initially, a $2D$ shape is represented by a configuration matrix X of size $k \times 2$, consisting of the Cartesian coordinates of the k landmarks outlining a $2D$ planar domain. However, as the object's shape encompasses all geometric information invariant to translations, rotations, and changes in scale (similarity transformations), we eliminate translations and scale changes from X by multiplying it with the Helmert submatrix, H [11], and normalizing it by its Frobenius norm. Thus, the pre-shape of the configuration matrix X is given by a $(k-1) \times 2$ matrix:

$$Z = \frac{HX}{||HX||} \tag{1}$$

The pre-shape space S_2^k, defined as the set of all possible pre-shapes, is a hypersphere of unit radius in $\mathbb{R}^{2(k-1)}$. Finally, to eliminate rotations and reflections, the $2D$ shape is obtained by taking the quotient of Z_X over all possible orthogonal transformations of the plane. Thus, the $2D$ shape space, Ω_2^k, is the quotient space of S_2^k under orthogonal transformations, i.e.,

$$\Omega_2^k = S_2^k / O(2). \tag{2}$$

Therefore, if π denotes the natural projection to the equivalence class, $\pi : S_2^k \to \Omega_2^k$, a shape $[X]_R$ is an orbit associated with the action of the orthogonal group $O(2)$ on the pre-shape; therefore, a shape is invariant under isometries in $\mathbb{R}^2$ [11].

On the other hand, given a configuration matrix X representing a shape $[X]_R$, another matrix belonging to the same equivalence class $[X]_R$ is given by $H^T Z$.

The space of reflection planar shapes, Ω_2^k, like the space of planar shapes, Σ_2^k, is a smooth manifold of dimension $2k-4$.

2.2. A Kernel Method in Ω_2^k

In [12], an embedding of the Kendall space Σ_2^k into a Euclidean space is defined. When reflection invariance is considered, this embedding leads to the following embedding, $\mathrm{i} : \Omega_2^k \to \mathbb{R}^{\frac{k(k-1)}{2}}$, from which we will introduce an extrinsic distance and a kernel method in Ω_2^k.

Theorem 1. *Let Z be a $(k-1) \times 2$ pre-shape of a given reflection shape $[X]_R$ and $v = \{v_1\, v_2\, \cdots\, v_{k-1}\}$ with $v_i \in \mathbb{R}^2$, $i \in \{1, \cdots, k-1\}$, the set of vectors defining Z.*

Then, the map is

$$\begin{aligned} \mathrm{i} : \Omega_2^k &\longrightarrow \mathbb{R}^{\frac{k(k-1)}{2}} \\ [X]_R &\longmapsto \mathrm{i}([X]_R) = \{< v_i, v_j >\}, \end{aligned} \tag{3}$$

where $i, j = 1, \ldots, k-1$, $i \leq j$ and $<, >$, which denotes the scalar product, is injective.

Proof. This is similar to the proof presented in [12] for the case of shapes in the Kendall space Σ_m^k, but notes that the generating set of polynomials invariant under the action of the group $O(n)$, unlike the group $SO(n)$, contains only scalar products and not determinants. □

The extrinsic distance between two shapes, $[X_1]$ and $[X_2]$, is given by the equation

$$d_e([X_1]_R, [X_2]_R) := \|\mathrm{i}([X_2]_R) - \mathrm{i}([X_1]_R)\|, \tag{4}$$

where $\| \cdot \|$ denotes the norm in $\mathbb{R}^{k(k-1)/2}$.

Finally, we consider the embedding of the Euclidean space $\mathbb{R}^{k(k-1)/2}$ into a reproducing kernel Hilbert space (RKHS) and work with the distance defined by the kernel, in addition to the Euclidean distance.

Proposition 1. *Let $K : \mathbb{R}^M \times \mathbb{R}^M \to \mathbb{R}$ be a positive definite kernel. Then,*

$$\widetilde{K} : \Omega_m^k \times \Omega_2^k \to \mathbb{R}, \quad ([X]_R, [Y]_R) \mapsto \widetilde{K}([x]_R, [Y]_R) := K(\mathrm{i}([X]_R), \mathrm{i}([Y]_R))$$

is a positive definite kernel on Ω_2^k.

In this paper, we use the Gaussian kernel $K : \Omega_2^k \times \Omega_2^k \longrightarrow \mathbb{R}$, defined as

$$K([X_1]_R, [X_2]_R) = \exp\Big(-\gamma d_e^2([X_1]_R, [X_2]_R)\Big). \tag{5}$$

The Gaussian kernels defined in the shape space Σ_2^k are the Procrustes Gaussian kernel [13] and the Projection Gaussian kernel [14], considering the 2D Kendall shape space Σ_2^k as a Grassmannian manifold.

2.3. An Extrinsic Mean Shape

In this section, we introduce a characterization of the submanifold $M = \mathrm{i}(\Omega_2^k)$ in $\mathbb{R}^{k(k-1)/2}$. Subsequently, we define the extrinsic mean as the projection onto M of the usual arithmetic mean in $\mathbb{R}^{k(k-1)/2}$.

Proposition 2. *Let $M \subset \mathbb{R}^{k(k-1)/2}$ be given by the vectors*

$$(a_i,\, b_{ij}) \in \mathbb{R}^{k(k-1)/2}, \tag{6}$$

where $i < j$ with $i, j \in \{1, 2, \cdots, k-1\}$, such that

1. $\sum_{i=1}^{k-1} a_i = 1$*, and* $a_i > 0$ *for* $i = 1, \ldots, k-1$.
2. $a_1\, b_{ij} = b_{1j} b_{1i} + \sqrt{(a_1 a_j - b_{1j}^2)(a_1 a_i - b_{1i}^2)}, \quad i = 2, \ldots, k-1, \quad i < j.$

Then $M = \mathrm{i}(\Omega_2^k)$.

Proof. Firstly, we observe that the dimension of M is $dim(M) = 2k-4$, therefore coinciding with the dimension of Ω_2^k. We need to demonstrate that $Im(\mathrm{i}) = M$.

Let

$$V = \left(x_i^2 + y_i^2, \quad x_i x_j + y_i y_j\right) = \left(\langle v_i, v_i\rangle, \langle v_i, v_j\rangle\right) \in Im(i). \tag{7}$$

Let us define $a_i = x_i^2 + y_i^2$. Then, $a_i \geq 0$, and since none of the Helmertized landmark coordinates coincide with the origin (Equation (3.7) of [11]), we have $a_i > 0$.

Moreover, since the pre-shape $Z \in S_2^k$, we have $\sum_{i=1}^{k-1} a_i = 1$.

Let

$$b_{ij} = \langle v_i, v_j\rangle = \langle (x_i, y_i)), (x_j, y_j)\rangle = ||v_i||\,||v_j|| \cos\theta_{ij}. \tag{8}$$

Since $\cos\theta_{ij} = \cos(\theta_{1j} - \theta_{1i})$, we obtain

$$b_{ij} = \sqrt{a_i}\sqrt{a_j}\cos(\theta_{1j} - \theta_{1i}).$$

Using the expression for the cosine of the difference of angles, we obtain

$$\begin{aligned} b_{ij} =& \sqrt{a_i}\sqrt{a_j}(\cos(\theta_{1j})\cos(\theta_{1i}) + \sin(\theta_{1j})\sin(\theta_{1i})) \\ =& \sqrt{a_i}\sqrt{a_j}\Big(\cos(\theta_{1j})\cos(\theta_{1i}) + \sqrt{(1-\cos^2(\theta_{1j}))(1-\cos^2(\theta_{1i}))}\Big). \end{aligned}$$

From Equation (8), we have

$$b_{ij} = \sqrt{a_i}\sqrt{a_j}\left(\frac{b_{1j}}{\sqrt{a_1}\sqrt{a_j}}\frac{b_{1i}}{\sqrt{a_1}\sqrt{a_i}} + \sqrt{\left(1 - \frac{b_{1j}^2}{a_1 a_j}\right)\left(1 - \frac{b_{1i}^2}{a_1 a_i}\right)}\right),$$

and then condition 2.

Therefore, $V \in M$, and we have demonstrated that $Im(\mathrm{i}) \subset M$. Now, let $V = \left(a_i, b_{ij}\right) \in M$, which satisfies conditions 1 and 2. We look for $[X] \in \Omega_2^k$ such that $\mathrm{i}([X]) = V$. We suppose that the pre-shape of X is defined as

$$Z = \begin{bmatrix} \sqrt{a_1} & 0 \\ \frac{b_{12}}{\sqrt{a_1}} & \frac{\sqrt{a_1 a_2 - b_{12}^2}}{\sqrt{a_1}} \\ \vdots & \vdots \\ \frac{b_{1k-1}}{\sqrt{a_1}} & \frac{\sqrt{a_1 a_{k-1} - b_{1k-1}^2}}{\sqrt{a_1}} \end{bmatrix}. \tag{9}$$

Then, $||Z|| = 1$ and $i([X]) = V$, so $M \subset Im(\mathrm{i})$. □

Next, we define a new extrinsic mean in space Ω_2^k.

Consider a sample of shapes, $[X_1]_R, \ldots, [X_n]_R \in \Omega_2^k$, and denote $\mathrm{i}([X_1]_R) = V_1, \ldots,$ $\mathrm{i}([X_n]_R) = V_n \in M \in \mathbb{R}^{k(k-1)/2}$. The Euclidean sample mean

$$\bar{V} = \frac{\sum_{i=1}^n V_i}{n} = \{\bar{a}_i, \bar{b}_{ij}\}$$

does not belong, in general, to M.

To define a sample mean shape in Ω_2^k, we first consider the following minimization problem:

Minimize $d_E^2(\bar{V}, M)$; that is, $\min_{W \in M} d_E^2(\bar{V}, W)$, where d_E denotes the Euclidean distance in $\mathbb{R}^{k(k-1)/2}$.

We aim to minimize the function

$$f(a_i, b_{ij}) = \sum_{i=1}^{k-1}(a_i - \bar{a}_i)^2 + \sum_{\substack{i=1 \\ i<j}}^{k-1}(b_{ij} - \bar{b}_{ij})^2,$$

subject to conditions 1 and 2 of Proposition 2.

Once we have obtained $W \in M$, which gives the minimum, the pre-shape Z of X is obtained from Equation (9), and the corresponding sample mean shape will be the shape $[X] \in \Omega_2^k$ such that $X = H^T Z$.

2.4. Image Database

Our approach was applied to the *erythrocytesIDB* image database (accessible at http://erythrocytesidb.uib.es). This dataset consists of images of peripheral blood smear samples obtained from patients with sickle cell disease in the Special Hematology Department of the General Hospital "Dr. Juan Bruno Zayas Alfonso" in Santiago de Cuba. The database provides the contours of 623 red blood cells, each represented by 295 points in $\mathbb{R}^2$, distributed approximately equidistantly along the boundary of each cell. A first-grade specialist in the Clinical Laboratory analyzed the images and classified them as circular, elongated, or cells with other deformations. We utilized the contours provided in the database, as optimizing the segmentation process was not the purpose of this study. Additional details about the sample preparation, image acquisition, segmentation, and other procedures are available on the image database homepage. The dataset contains 202 images of normal cells, 210 of sickle cells, and 211 of cells with various other deformations. Figure 1 presents three sample images from the database. It is important to note that the class of erythrocytes with other deformations includes cells with minor irregularities and shapes resembling both normal and sickle cells, which could lead to misclassification and reduced interpretability. Deformations of red blood cells are associated with several significant diseases. In some of these conditions, such as thalassemias, the presence of microcytic (small) and hypochromic erythrocytes is a central characteristic, significantly impairing oxygen transport. Therefore, in these diseases, the size of the erythrocytes plays a crucial role. However, in the case of sickle cell anemia, it is the sickle-shaped form of the erythrocytes that primarily impacts the disease's progression.

The contours of all 623 cells in the database were uniformly characterized by 295 points in $\mathbb{R}^2$. To determine the major axis of each cell, we calculated the maximum distance among all pairs of the 295 points defining the contour of each cell. Among the two points defining this major axis, the one with the largest x-coordinate was chosen as the first landmark, ensuring a consistent reference point for all cells. Using this initial landmark as the starting reference, the remaining points were treated as pseudo-landmarks, thereby establishing correspondence between the landmarks of all cells. This approach, combined with the distinctive shapes of normal and sickle cells and their consistent representation, allowed each cell to be regarded as a shape in the Kendall reflection shape space Ω_2^k.

As we will discuss in Section 3, we performed a comparison between the classification results obtained using 295 landmarks per cell and those obtained using $295/5 = 59$ landmarks per cell (i.e., once the first landmark was selected from the major axis of the cell, we only considered 59 equidistant points from this landmark, disregarding the rest). Since the difference in the classification results was not significant, we used $k = 59$ landmarks in the classification process.

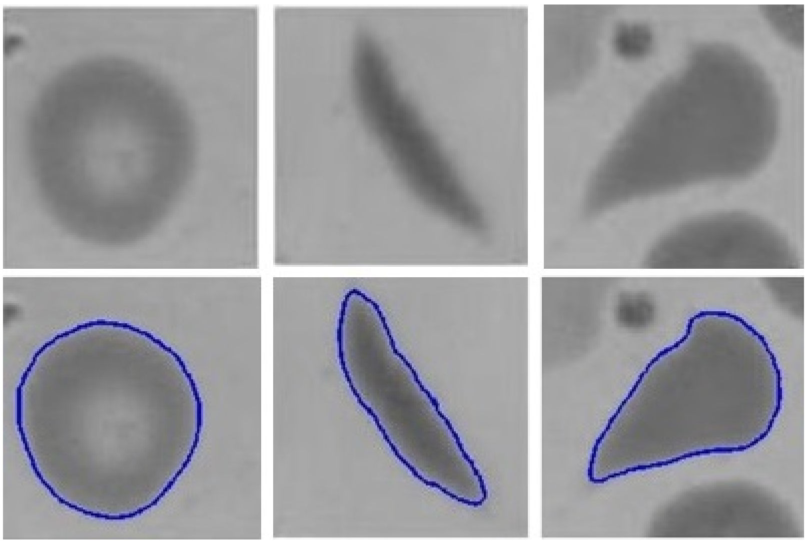


Figure 1. Examples of erythrocytes categorized as normal (**left**), sickle (**center**), and exhibiting other deformations (**right**). **Top**: the original cell; **bottom**: the perimeter of segmented regions (depicted in blue).

2.5. Supervised Classification

To perform the supervised classification, we used the default parameters provided by the `R` package, as our primary goal was not to optimize the models but rather to treat erythrocytes as shapes in Kendall's shape space and, for the first time, apply machine learning classification models by embedding Kendall's space into a Euclidean space. In this context, we did not use a validation dataset. Consistent with other studies using the same dataset, the set of 623 cells was randomly divided such that 80% of the cells in each class were used for training and 20% for testing. Consequently, the test set consisted of 40 normal cells, 42 sickle cells, and 42 cells with other deformations.

Although in this paper we used the default parameters for each classification method, we performed a 5-fold cross-validation process to ensure the reliability of the results. Additionally, for the SVM method, we repeated the classification by modifying the parameters to further strengthen the robustness of the results.

For the classification, we considered the four machine learning algorithms outlined below. For all algorithms, we calculated the pre-shapes:

$$Z_i = \frac{HX_i}{\|HX_i\|},$$

where the Helmert submatrix is obtained using the `R` command `helm(k)`. Next, we obtained the vectors $\mathrm{i}([X_i]_R)$ in $\mathbb{R}^{k(k-1)/2}$ from Equation (3).

2.5.1. Algorithm 1: k-Nearest Neighbors (k-NN)

k-Nearest Neighbors (k-NN) is a simple, non-parametric supervised learning algorithm used for classification tasks. It works by identifying the k-nearest data shapes (neighbors) to a given observation based on the Euclidean distance. The algorithm assigns a class label to the observation based on the majority class of its nearest neighbors. In our case, the algorithm looked at the 5 closest neighbors. However, in Section 3, we will justify this choice by presenting a graph of the error rate, i.e., the proportion of incorrect predictions, for different values of the number of neighbors k.

2.5.2. Algorithm 2: Support Vector Machine (SVM)

The Support Vector Machine (SVM) is a supervised learning algorithm used for classification tasks. It works by finding the hyperplane that best separates different classes in the feature space. The goal is to maximize the margin between the closest shapes of the classes (support vectors) and the hyperplane. In our case, we considered the SVM function in `R` (from the `e1071` package) with the radial basis function (RBF) kernel and the default parameters of the cost ($C = 1$) and $\gamma = 0.0005844535$. These parameter values suggest that the model aims for a balance between generalization and fitting the training data. In the Results Section, Section 3, we will analyze how variations in these parameters have little effect on the classification accuracy, as long as γ remains within certain limits.

2.5.3. Algorithm 3: Naive Bayes

Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem, which assumes that the features used for classification are independent of each other given the class label. The model calculates the probability of each class given the input features and assigns the class with the highest probability to the observation. The default parameter values used in the model are as follows:

Laplace Smoothing (laplace): This parameter, set to zero by default, controls the degree of smoothing applied to avoid zero probabilities.

Kernel Usage (usekernel): Set to FALSE by default, this parameter determines whether kernel-based estimations are used for the distributions of continuous variables. When set to FALSE, it assumes the variables follow normal (Gaussian) distributions.

The Distribution Weighting Factor (fL): Set to 1 by default, this parameter affects the weight of observed data in the estimation of distributions.

2.5.4. Algorithm 4: Random Forest

Random Forest is an ensemble learning method primarily used for classification tasks. It works by constructing a multitude of decision trees during training and outputting the mode of the individual trees for classification. In our case, 100 trees were grown in the forest to ensure a robust model. We used the `R` library randomForest (`library(randomForest)`). The remaining default values used in the model are as follows: The number of variables per split (mtry), set to 41: this parameter determines how many variables are considered when searching for the best split at each tree node. The minimum node size (nodesize), set to 1: this parameter specifies the minimum number of observations required in a node for it to be eligible for further splitting. Finally, the maximum number of nodes (maxnodes) was set to NULL (default); this parameter defines an upper limit on the total number of nodes a tree can have. When left as NULL (the default value), no explicit limit is imposed.

2.6. Unsupervised Clustering

Since we had an embedding of the space Ω_2^k into a Euclidean space, based on the distances defined in Section 2.2, we proposed using three k-means algorithms to perform the unsupervised clustering of the cells into three groups: normal cells, sickle cells, and cells with other deformations. Therefore, the number of clusters considered in the algorithms was $k = 3$.

2.6.1. Algorithm 1: Extrinsic Distance

In the first algorithm, once we had all the cells characterized by the landmarks defined based on their diameter, we calculated the pre-shapes Z_i and we obtained the vectors $\mathrm{i}([X_i]_R)$ in $\mathbb{R}^{k(k-1)/2}$ from Equation (3). Finally, we used the extrinsic distance defined in Equation (4) and applied the standard `R` functions for k-means clusters (`kmeans`) [15].

Note that in this algorithm, we used the Euclidean sample mean $\bar{V} = \frac{\sum_{i=1}^{n} V_i}{n}$ instead of its projection onto the manifold M.

2.6.2. Algorithm 2: Classification Using Known Templates

In the second algorithm, since normal cells approximate a circle and elongated cells an ellipse, we calculated the extrinsic distances considered in the second algorithm from the 623 cells to a circle and to an ellipse whose major axis was three times the minor axis. We then associated each cell with these two values and performed k-means clustering.

In the two previous algorithms, the parameters used in the k-means method included, in addition to the number of clusters set to $k = 3$, the initialization parameter (`nstart`), which was set to its default value of 25. This parameter specifies how many times the algorithm initializes with different configurations of centroids. Additionally, the maximum number of iterations was also set to its default value of 100. This parameter defines the maximum number of iterations the algorithm performs to adjust the centroids and assign data points to clusters.

2.6.3. Algorithm 3: Kernel k-Means

In the third algorithm, we considered the embedding of the shape space into a reproducing kernel Hilbert space (RKHS) [12], and we operated with the distance defined by the kernel instead of the Euclidean distance. This approach offered several advantages, including the enhanced separability of classes in the transformed feature space, thereby facilitating the classification task. Additionally, it allowed for flexibility in kernel selection, although we specifically proposed the default kernel (`rbfdot`), which corresponds to a Gaussian kernel, with the kernel parameter γ set to 8 in Equation (5).

Therefore, once we obtained the vectors $\mathrm{i}([X_i]_R)$ in $\mathbb{R}^{k(k-1)/2}$, we applied the `kkmeans` function of the `kernlab` package in `R`.

The rest of the parameters used in this case were the number of clusters set to 3, initialization (`nstart`) set to 1, and the number of iterations set to 100.

All `R` scripts used in this work are available upon request.

3. Results

First, we will verify that the results obtained when using 295 landmarks along the contour of each cell (Table 1) and those obtained with 59 landmarks (Table 2) are similar. This will be demonstrated through the representation of the confusion matrix for the three classes $\{N, S, OD\}$ and the accuracy for each supervised classification algorithm. The accuracy results in both tables highlight that the Random Forest method is the most sensitive to a high number of landmarks, as being a tree-based model, it increases the complexity and the likelihood of overfitting. Therefore, given the reduced computational cost associated with using 59 landmarks, we will proceed with this smaller number of landmarks for subsequent analyses.

Table 1. Confusion matrix and measures for supervised classification with $k = 3$ and 295 landmarks.

	Alg. 1			**Alg. 2**			**Alg. 3**			**Alg. 4**		
	N	***S***	***OD***	***N***	***S***	***OD***	***N***	***S***	***OD***	***N***	***S***	***OD***
$\widehat{N}$	40	0	10	39	0	2	39	0	3	38	0	3
$\widehat{S}$	0	42	1	0	40	0	0	42	1	0	42	1
$\widehat{OD}$	0	0	31	1	2	40	1	0	38	2	0	36
Acc		91.12			95.97			95.97			95.16	

The metrics used to evaluate the results were the sensitivity or True Positive Rate (TPR), the precision (P), the specificity or True Negative Rate (TNR), and the F1 score (F1). The TPR is the number of correct positive predictions for each class divided by the total number of objects in that class. The precision is the number of objects correctly classified in a class divided by the total number of instances classified as positive in that class. The TNR is the number of correct negative predictions for each class divided by the total number of objects that do not belong to that class. The F1 score is the harmonic mean of the precision and sensitivity, providing a balanced measure between both metrics.

Moreover, the accuracy metric (Acc) measures the proportion of correct predictions made by the model out of the total number of predictions. The SDS score, introduced in [16], is calculated as the ratio of the sum of true positives across the three classes and the number of sickle cells incorrectly classified as other deformations (and vice versa) to the sum of this numerator and the total number of incorrect classifications involving normal cells. This score serves as an indicator of the method's effectiveness in supporting the analysis of the studied disease.

To justify the number of neighbors selected for applying the k-NN algorithm, in Figure 2, we present the error rate (the proportion of cells in the dataset that were misclassified) for different values of the number of neighbors. As observed, for values greater than five, the error increases.

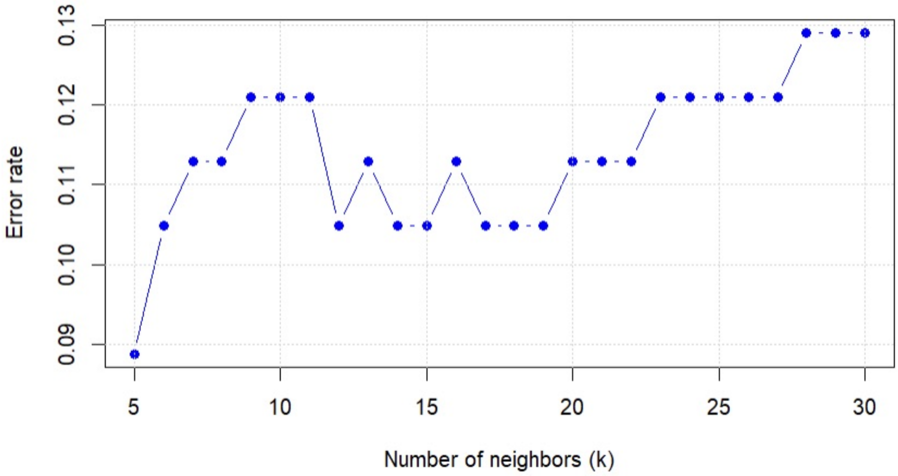


Figure 2. Error rate as a function of the number of neighbors selected in Algorithm 1.

When analyzing the results from the supervised classification algorithms (Table 2), we observe that Alg. 1 (k-NN) and Alg. 2 (the SVM) both demonstrate strong performance in distinguishing between normal cells and sickle cells, with no misclassifications between these two classes. However, a few cells with other deformities (ODs) were misclassified as either normal or sickle cells, particularly with k-NN, while the SVM shows a notable reduction in these misclassifications. In the SVM method, we performed the classification by varying the parameters C and γ to analyze how these changes affected the classification. As we can see in Figure 3, for values of C between 0.01 and 100 and γ between 0.0001 and 0.005, the accuracy of the classification remains above 90%. The fact that the accuracy stays consistent despite the variation in these two parameters reinforces the robustness of the model and its ability to generalize.

Table 2. Confusion matrix and measures for supervised classification with $k = 3$ and 59 landmarks.

	Alg. 1			**Alg. 2**			**Alg. 3**			**Alg. 4**		
	N	***S***	***OD***	***N***	***S***	***OD***	***N***	***S***	***OD***	***N***	***S***	***OD***
$\widehat{N}$	40	0	9	40	0	3	39	0	3	38	0	1
$\widehat{S}$	0	42	1	0	42	2	0	41	1	0	42	0
$\widehat{OD}$	0	0	32	0	0	37	1	1	38	2	0	41
TPR	100	100	76.19	100	100	88.10	97.50	97.62	90.48	95	100	97.62
TNR	89.29	98.78	100	96.43	97.56	100	96.43	98.78	97.56	98.81	100	97.56
P	81.63	97.67	100	93.02	95.45	100	92.86	97.62	95	97.44	100	95.35
F_1	89.89	98.82	86.49	96.38	97.67	93.67	95.12	97.62	92.68	96.20	100	96.47
Acc/SDS	91.94/92.74			95.97/97.58			95.16/96.77			97.58/97.58		

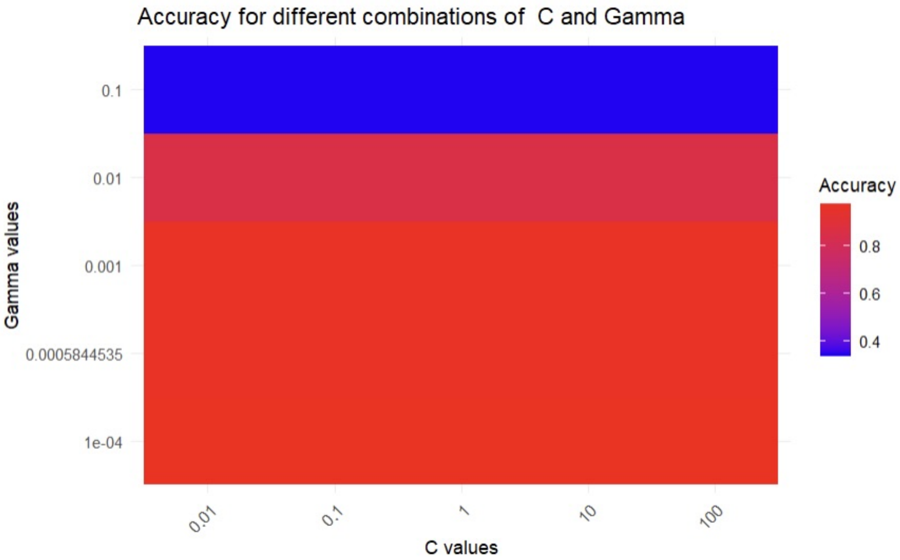


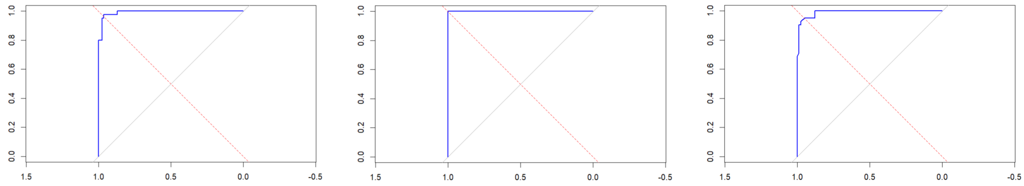
Figure 3. Classification accuracy with the SVM method when varying the parameters.

Alg. 3 (Naive Bayes) and Alg. 4 (Random Forest), however, provide superior sensitivity (the True Positive Rate) for cells with other deformities. Alg. 4 (Random Forest), in particular, excels across the board, achieving the highest overall accuracy (97.58%) and maintaining a perfect balance between sensitivity and specificity across all classes, as indicated by its SDS score (97.58). This algorithm also reduces misclassification errors for cells with other deformities, showing the highest TPR (97.62%) for this challenging class and achieving perfect precision (100%) for sickle cells.

In summary, Alg. 4 (Random Forest) offers the best overall performance, combining high accuracy, excellent sensitivity, and balanced performance across all cell types. It outperforms the other methods with the default search parameters, particularly in handling cells with other deformities, making it the most reliable choice for supervised classification in this context. Alg. 2 (the SVM) also shows strong results, especially for normal and sickle cells, but does not match the overall effectiveness of Random Forest.

To evaluate the proposed classification models, we utilized two important tools: the ROC (Receiver Operating Characteristic) curve and the AUC (Area Under the Curve) score. These metrics provide insight into the separability of the classes across all possible thresholds, effectively showing how well the model distinguishes each class. For each group, we calculated a ROC curve and an AUC score by considering that group as the positive class and all other groups as the negative class (a "One-vs-Rest" approach).

While we computed the ROC curves for all four algorithms used in supervised classification, Figure 4 displays only the ROC curves for the k-NN method, as the curves for the other methods were very similar. As seen in the results, the ROC curves for all three classes exhibit a very good fit. Moreover, the AUC scores for the three groups across all four classification methods were consistently above 0.99. These results indicate that all three models can be considered excellent, as an AUC value near one signifies a high level of class separability and strong classification performance.



Normal group Sickle group Other deformations group

Figure 4. ROC curves for the k-NN method.

When comparing the results of our methods with those obtained using other approaches in the literature, such as shape descriptors like the Circular Shape Factor (CSF) or features derived from Fourier methods (see Table 3 in [7]), we find that our methods generally achieve better percentages across all considered metrics. Moreover, when comparing our results to those obtained using parametric curve representations of shapes [8], it is notable that the classification of cells in those studies only considered k-NN algorithms. In contrast, by embedding shapes into a Euclidean space, our approach allowed us to employ a variety of classification algorithms. Comparing Tables 1 and 2 in [8] with Table 2 in our work, we observe that the classification results obtained with our metrics are generally superior. For instance, the maximum overall accuracy they report is 94.2%, while with Algorithms 2, 3, and 4, we achieve an accuracy clearly exceeding 95%.

Results of Unsupervised Clustering

In this section, $\{N, S, OD\}$ represent the same three classes as in the preceding section, and we used the same metrics to develop the unsupervised clustering. The confusion matrices corresponding to the four algorithms are presented in Table 3.

Table 3. Confusion matrix and measures for unsupervised clustering with $k = 3$.

	Alg. 1			**Alg. 2**			**Alg. 3**		
	N	S	OD	N	S	OD	N	S	OD
$\widehat{N}$	200	0	77	201	0	77	199	0	59
$\widehat{S}$	0	202	7	0	203	4	0	206	11
$\widehat{OD}$	2	8	127	1	7	130	3	4	141
TPR	99.01	96.19	60.19	99.50	96.67	61.61	98.51	98.10	66.82
TNR	81.71	98.30	97.57	81.71	99.03	98.06	85.85	97.34	98.30
P	72.20	96.65	92.70	72.30	98.07	94.20	77.13	94.93	95.27
F_1	83.51	96.42	72.99	83.75	97.38	74.50	86.52	96.49	78.55
Acc/SDS	84.91/87.32			85.71/87.48			87.64/90.05		

Table 3 presents the performance of three unsupervised clustering algorithms (Alg. 1, Alg. 2, and Alg. 3) evaluated on the dataset. The analysis reveals that Alg. 3, which uses kernel k-means, consistently outperforms the others, particularly in classifying cells with other deformities (ODs). Alg. 3 achieves the highest overall accuracy (87.64%) and excels in key metrics like the precision and F1 score across all classes, demonstrating its robustness and effectiveness in this task.

Alg. 2 also performs well, particularly in classifying normal and sickle cells, with slightly lower performance in the OD class. Alg. 1 lags behind the other two, with lower accuracy and precision, especially for OD cells, though it still provides reasonable results for normal and sickle cells. Overall, Alg. 3 is the most reliable and balanced choice for accurate classification across all cell types.

When cells are considered as closed parameterized curves in the shape space, accounting for their parameterization, the confusion matrices found in Table 3 of [8] apply. In that study, because working with intrinsic Fréchet means in the space of parameterized curves is complex, unsupervised clustering was performed using the Partitioning Around Medoids (PAM) method.

In general, the metrics of our three methods yield better results for normal and sickle cells. However, for cells with other deformities, the methods tend to misclassify some as sickle and especially as normal cells. While this is noteworthy, since cells with other deformities could indicate other hematological pathologies, the primary goal of our classification is to detect the presence and percentage of sickle cells. Thus, the most critical error is the misclassification of sickle cells, whether by underestimating or overestimating their percentage. In this regard, as shown in Table 3, our metrics indicate excellent performance in accurately classifying sickle cells.

4. Discussion

There are numerous approaches in the literature for classifying red blood cells into three categories: healthy, sickle cells, and those with other deformities. Recently, one approach has involved considering cell boundaries as parameterized curves and working in a shape space that is invariant under rigid motions and reparameterizations. In this paper, we present a new approach that avoids the complexities of reparameterization by defining mathematical landmarks along the cell boundaries. Each cell's shape is then identified as a point in the differentiable manifold known as the Kendall space of reflection shapes. To facilitate the application of machine learning classification methods, we proposed an embedding of the Kendall space into a Euclidean space. This new classification methodology was applied to a database of red blood cells, and the results from various algorithms show excellent classification performance. Specifically, in supervised classification, the Support Vector Machine (SVM) and Random Forest algorithms achieved overall accuracy rates of 95.97% and 97.58%, respectively, with both obtaining a DS score of 97.58%. In the case of unsupervised clustering, the kernel k-means method produced the best results.

When comparing our results with existing approaches in the literature, nearly all the metrics are comparable to, if not superior to, those previously reported. A significant advantage of our method is the use of a Euclidean space with only 59 landmarks, as opposed to 295 points defining the curve boundaries, which greatly reduces the computational complexity.

In Table 4, we compare the accuracy obtained in the supervised classification using our methods to the accuracy reported in other studies that also classify red blood cells into three groups. These comparisons are based on the same dataset and utilize the same data split of 80% for training and 20% for testing.

Previous studies in this domain have employed various approaches. For instance, Ref. [7] focuses on the planar shape space of parameterized curves, using a distance invariant only under arc-length parameterizations, while Ref. [5] applies the classical k-Nearest Neighbor (k-NN) technique with contour descriptors derived from integral geometry methods. Similarly, Ref. [17] evaluates three widely used supervised classifiers—Naive Bayes, k-NN, and the Support Vector Machine (SVM)—based on a set of nine numerical shape descriptors. Scenario 1 in [9] explores lightweight models with variations in layers and filters, using the original *erythrocytesIDB* image database as the input. Notably, the best accuracy in this scenario was achieved with Method 2, incorporating a multiclass SVM classifier. Furthermore, Ref. [18] identifies optimal classification techniques and features for cell morphology analysis, while Ref. [8] investigates the planar shape space of parameterized curves using the elastic metric derived from the square root velocity function.

As shown in Table 4, the accuracy achieved with our methods is comparable to, and often exceeds, the performance reported in these studies. This highlights the robustness and effectiveness of our approach.

This study presents a new approach by representing red blood cell shapes as points in the Kendall space of reflection shapes and characterizing this space as a subset of a Euclidean space. This representation allows for the application of machine learning classification methods. These methods yield very remarkable classification results compared to other approaches using the same red blood cell database. However, the proposed methodology also has certain limitations.

This study primarily focuses on the shape of red blood cells. While this is sufficient for sickle cell anemia, where the detection of sickle cells primarily depends on their distinctive shape, for other diseases, deformations or changes in the cell size may play a key role. Furthermore, the determination of landmarks at the boundaries of red blood cells is based on the cell diameter. Although this approach is effective for red blood cells, given the particular geometry of healthy cells and those affected by sickle cell anemia, it may not be suitable for applications in other fields where shapes have different characteristics. These limitations may certainly lead to new approaches for future work.

Table 4. Comparison of accuracy results with previous methods employed on the erythrocytesIDB dataset.

Method	Accuracy (%)
Gual-Arnau et al. (2015) *Image Anal Stereol.* [7]	93.42
Gual-Arnau et al. (2015) *Med. Biol. Eng. Comput.* [5]	96.10
Rodrigues et al. (2016) *Workshop de Visao Computacional.* [17]	94.59
De Faria et al. (2018) *Workshop de Visao Computacional.* [19]	93.67
Alzubaidi et al. (2020) *Electronics.* (Scenario 1) [9]	90.80
Petrović et al. (2020) *Comput. Biol. Med.* [18]	95.20
Epifanio et al. (2020) *Image Anal Stereol.* [8]	95.10
Our methods (Table 2)	91.94-95.97-95.16-97.58

Author Contributions: Conceptualization, X.G.-A. and L.G.-V.; methodology, X.G.-A. and L.G.-V.; software, L.G.-V.; validation, X.G.-A. and L.G.-V.; formal analysis, X.G.-A. and L.G.-V.; investigation, X.G.-A. and L.G.-V.; resources, X.G.-A. and L.G.-V.; data curation, L.G.-V.; writing—original draft preparation, X.G.-A.; writing—review and editing, X.G.-A.; visualization, X.G.-A. and L.G.-V.; supervision, X.G.-A.; project administration, X.G.-A.; funding acquisition, X.G.-A. All authors have read and agreed to the published version of this manuscript.

Funding: The first author has been partially supported by the Spanish Ministry of Science and Innovation (PID2020-115930GA-100 and PID2022-141699NB-I00).

Data Availability Statement: Data is contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Taherisadr, M.; Nasirzonouzi, M.; Baradaran, B.; Mehdizade, A. New approach to red blood cell classification using morphological image processing. *Shiraz E-Med. J.* **2013**, *14*, 44–53.
2. Gual-Vayà, L. Classification of Red Blood Cells From a Geometric Morphometric Study. *Image Anal. Stereol.* **2024**, *43*, 109–119. [CrossRef]
3. Bronkorsta, P.; Reinders, M.; Hendriks, E.; Grimbergen, J.; Heethaar, R.; Brakenhoff, G. On-line detection of red blood cell shape using deformable templates. *Pattern Recogn. Lett.* **2000**, *21*, 413–424. [CrossRef]
4. Lee, H.; Chen, Y.P. Cell morphology based classification for red cells in blood smear images. *Pattern Recogn. Lett.* **2014**, *49*, 155–161. [CrossRef]
5. Gual-Arnau, X.; Herold-García, S.; Simó, A. Erythrocyte shape classification using integral-geometry-based methods. *Med. Biol. Eng. Comput.* **2015**, *53*, 623–633. [CrossRef] [PubMed]
6. Navya, K.; Prasad, K.; Singh, B. Analysis of red blood cells from peripheral blood smear images for anemia detection: A methodological review. *Med. Biol. Eng. Comput.* **2022**, *60*, 2445–2462.
7. Gual-Arnau, X.; Herold García, S.; Simó, A. Geometric Analysis of Planar Shapes with Applications to Cell Deformations. *Image Anal. Stereol.* **2015**, *34*, 171–182. [CrossRef]
8. Epifanio, I.; Gual-Arnau, X.; Herold-Garcia, S. Morphologycal Analysis of Cells by Means of an Elastic Metric in the Shape Space. *Image Anal. Stereol.* **2020**, *39*, 281–291.
9. Alzubaidi, L.; Fadhel, M.A.; Al-Shamma, O.; Zhang, J.; Duan, Y. Deep learning models for classification of red blood cells in microscopy images to aid in sickle cell anemia diagnosis. *Electronics* **2020**, *9*, 427. [CrossRef]
10. Jennifer, S.; Shamim, M.; Reza, A.; Siddique, N. Sickle cell disease classification using deep learning. *Heliyon* **2023**, *9*, e22203. [CrossRef] [PubMed]
11. Dryden, I.; Mardia, K.V. *Statistical Shape Analysis: With Applications in R*; John Wiley and Sons: Hoboken, NJ, USA , 2016.
12. Gimeno, V.; Gual-Arnau, X.; Ibáñez, M.; Simó, A. A Gaussian kernel for Kendall's space of m-D shapes. *Pattern Recognit.* **2023**, *144*, 109887 .
13. Jayasumana, S.; Salzmann, M.; Li, H.; Harandi, M. A framework for shape analysis via Hilbert space embedding. In Proceedings of the 2013 IEEE International Conference on Computer Vision (ICCV), Sydney, Australia, 1–8 December 2013; pp. 1249–1256.
14. Jayasumana, S.; Hartley, R.; Salzmann, M.; Li, H.; Harandi, M. Kernel methods on Riemannian manifolds with Gaussian RBF kernels. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 2464–2477. [CrossRef] [PubMed]
15. Kassambara, A. *Practical Guide to Cluster Analysis in R*; STHDA. 2017. Available online: https://xsliulab.github.io/Workshop/2021/week10/r-cluster-book.pdf (accessed on 1 October 2024).

16. Delgado-Font, W.; Escobedo-Nicot, M.; Gonzalez-Hidalgo, M.; Herold-García, S.; Jaume-i-Capó, A. Diagnosis support os sickle cell anemia by classifying red blood cell shape in peripheral blood images. *Med. Biol. Eng. Comput.* **2020**, *6*, 1265–1284. [CrossRef] [PubMed]
17. Rodrigues, L.F.; Naldi, M.C.; Mari, J.F. Morphological analysis and classification of erythrocytes in microscopy images. In Proceedings of the 2016 Workshop de Visao Computacional, Campo Grande, Brazil, 9–11 November 2016; pp. 9–11.
18. Petrović, N.; Moyà-Alcover, G.; Jaume-i-Capó, A.; González-Hidalgo, M. 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. *Comput. Biol. Med.* **2020**, *126*, 104027. [CrossRef] [PubMed]
19. de Faria, L.C.; Rodrigues, L.F.; Mari, J.F. Cell classification using handcrafted features and bag of visual words. In Proceedings of the Workshop de Visão Computacional, Ilhéus, BA, Brazil, 2018; pp. 68–73. Available online: https://www.researchgate.net/publication/331181776_Cell_classification_using_handcrafted_features_and_bag_of_visual_words (accessed on 1 October 2024).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.