PLOS ONE

RESEARCH ARTICLE

Identification of the selected soil bacteria genera based on their geometric and dispersion features

Aleksandra Konopka[1]*, Ryszard Kozera[1,2], Lidia Sas-Paszt[3], Pawel Trzcinski[3], Anna Lisek[3]

1 Institute of Information Technology, Warsaw University of Life Sciences - SGGW, Warsaw, Poland, **2** School of Physics, Mathematics and Computing, The University of Western Australia, Perth, Australia, **3** Department of Microbiology and Rizosphere, The National Institute of Horticultural Research, Skierniewice, Poland

* aleksandra_konopka@sggw.edu.pl

Check for updates

OPEN ACCESS

Citation: Konopka A, Kozera R, Sas-Paszt L, Trzcinski P, Lisek A (2023) Identification of the selected soil bacteria genera based on their geometric and dispersion features. PLoS ONE 18(10): e0293362. https://doi.org/10.1371/journal.pone.0293362

Editor: Carlos Fernandez-Lozano, University of A Coruña, SPAIN

Received: April 25, 2023

Accepted: October 10, 2023

Published: October 27, 2023

Copyright: © 2023 Konopka et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The microscopic image dataset and the code written in support of this publication is publicly available at 10.5281/zenodo.7789436.

Funding: This research was supported by The National Centre for Research and Development within the framework of the project BIOSTRATEG, grant number BIOSTRATEG3/344433/16/NCBR/2018. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

The visual analysis of microscopic images is often used for soil bacteria recognition in microbiology. Such task can be automated with the aid of machine learning and digital image processing techniques. The best results for soil microorganism identification usually rely on extracting features based on color. However, accommodating in the latter an extra impact of lighting conditions or sample's preparation on classification accuracy is often omitted. In contrast, this research examines features which are insensitive to the above two factors by focusing rather on bacteria shape and their specific group dispersion. In doing so, the calculation of layout features resorts to *k*-means and mean shift methods. Additionally, the dependencies between specific distances determined from bacteria cells and the curvature of interpolated bacteria boundary are computed to extract vital geometric shape information. The proposed bacteria recognition tool involves testing four different classification methods for which the parameters are iteratively adjusted. The results obtained here for five selected soil bacteria genera: *Enterobacter*, *Rhizobium*, *Pantoea*, *Bradyrhizobium* and *Pseudomonas* reach 85.14% classification accuracy upon combining both geometric and dispersion features. The latter forms a promising result as a substitutive tool for color-based feature classification.

Introduction

Identification of bacteria can be realized with the use of many molecular techniques, including ribotyping, repetitive extragenic palindromic PCR (Rep-PCR), denaturing gradient gel electrophoresis (DGGE), terminal (T)-restriction fragment length polymorphism (T-RFLP), multilocus sequence typing (MLST) and whole-genome sequencing (WGS) [1]. MLST uses DNA sequencing of internal fragments of the housekeeping gene loci (seven in number) of bacterial strains to characterize alleles [2]. In practice, a common stance for bacteria identification is based on sequence analysis of 16 SrRNA gene [3, 4] and MLST unveiling the same intraspecific genetic structure patterns as genomes [5]. In bacteria recognition process, the morphological

Competing interests: The authors have declared that no competing interests exist.

features can also be considered while analyzing the microscopic images. However, sometimes it is hard to distinguish between different bacteria species due to their morphological similarities within a genera [6]. The image-based identification can be tedious and laborious.

The aim of this research is to create a system that automates the process of microscopic image classification. Incorporating the computerized approach facilitates the identification process replacing or supporting human expertise and eyesight assessment with the modern computer vision image processing techniques. Machine learning methods used in this paper have already been applied to solve pattern recognition, prediction and classification problems in various fields of biology [7] and, in particular, to identify the microorganisms [8]. Some bacteria can be easily discerned from others due to their specific morphological features e.g. *Mycobacterium tuberculosis* [9] and *Escherichia Coli* [10] both having characteristic shapes. Here a fast and robust recognition scheme is in demand as these bacteria may inflict serious human illnesses. Some works perform classification not on the genera or species level but defining each class as a shape type [11]. The features relying on shape, texture or on pixel-based measures are applied in bacteria classification [12–15]. In this paper, the classification task is accomplished on the genera level via differentiating microscopic images of five selected soil bacteria genera: *Enterobacter*, *Rhizobium*, *Pantoea*, *Bradyrhizobium* and *Pseudomonas* (see Fig 1) grown in specific conditions on selected medium. Some of these bacteria genera have a positive impact on plant growth while the others are pathogenic. For this reason it is important to accurately classify their character [16, 17].

Identification of microorganisms with machine learning methods is widely applied for recognition of pathogens causing human infections (see e.g. [12]). In contrast, the topic of soil microorganism classification has not been so-far extensively investigated. In case of image-based soil microorganism identification discussed in [18], the analysis of color features used for bacteria recognition yields up to 97% of classification accuracy (ACC). In the latter work, the goal was to create a system enabling automatic recognition of samples that are preprocessed by the microbiologists. The introduced chemical reactions result in the color change of samples depending on the species of the microorganism which ultimately facilitates the efficiency of the classifier in achieving more accurate results.

In our research a different approach is adopted. The microscopic images can be taken with various microscopes and under different lighting conditions. In addition, the photographed samples can also be processed by the microbiologists upon administering a contrast or initiating a chemical reaction. Furthermore, the analyzed samples are usually colored with dyes to improve visibility of the objects examined under the microscope. In order not to rely on these factors, the different types of features based on bacteria geometry and their group dispersion are considered in this work which yields an alternative for the color-based traits classification. Developing such a set of features can help to create an automatic program that performs an accurate classification on both raw images and on those subjected to chemical reactions. The computations are performed here on images of bacteria samples that are not earlier processed by the microbiologists. In the prior research, the combination of geometric and texture features [19] calculated on the same image dataset resulted in up to 97% classification accuracy. However, in this research features based on texture are excluded as they rely on luminance (i.e. pixels intensity values which in turn may depend on lighting conditions). Instead, only features related directly to the geometry and dispersion of the analyzed objects are considered. The highest classification accuracy obtained here for such a set of features equals 85.14%. The present findings suggest that alternative feature types have the potential to supplant chrominance and luminance features in the realm of bacterial classification. Such an approach would enable classification with comparable precision for images captured under diverse illumination conditions, amalgamating preprocessed and raw images, as the outcome remains impervious to

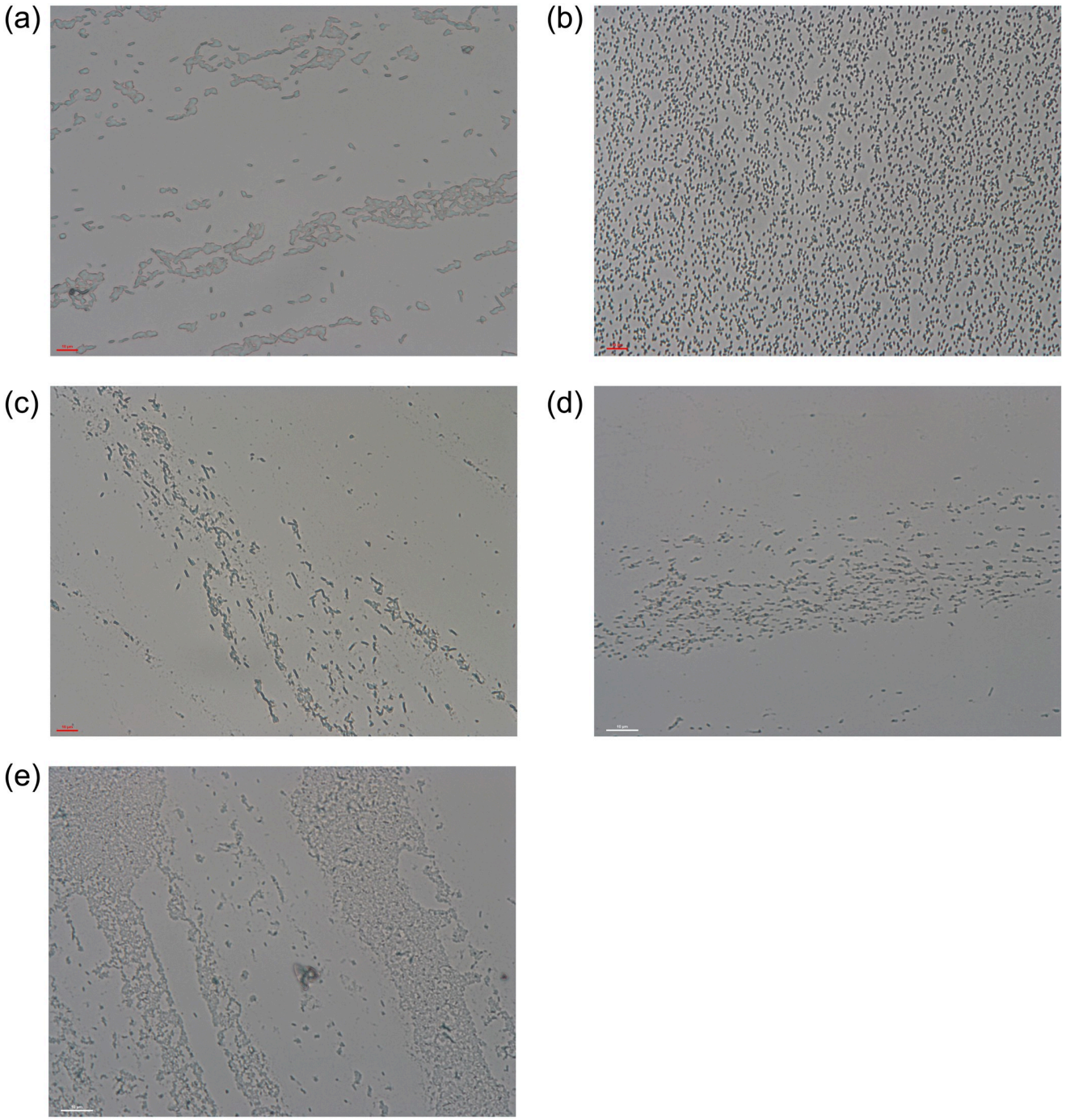

Fig 1. Examples of bacteria images: (a) *Enterobacter*, (b) *Rhizobium*, (c) *Pantoea*, (d) *Bradyrhizobium* and (e) *Pseudomonas*. For more pictures see URL link: bit.ly/3TwOgFB.

https://doi.org/10.1371/journal.pone.0293362.g001

color and light influences. Notably, the investigated set of microscopic images demonstrated an accuracy of 95.6% for exemplary color features, indicating a remaining deviation of 10% in classification accuracy. Nonetheless, given the multifaceted nature of this issue, further exploration of various factors is warranted, and the classification results are anticipated to be

enhanced upon adjustment of pertinent variables as explicated in the ensuing discussion section.

Material

Strains X58AD (*Pantoea sp*), Pi72ED (*Enterobacter sp*), Ps118AA (*Pseudomonas sp*) were grown for 48 hours in 26 Celsius degrees on Plate Count Agar (BTL P-0037) medium. E77AO (*Rhizobium*) bacteria strain and a strain that was not present in Symbio-Bank (*Bradyrhizobium*) were grown on Yeast Mannitol Agar medium for 96 hours in 26 Celsius degrees. Bacteria of each strain were collected from a single colony and transferred on the surface of glass plate. In the next step, a drop of sterile water was added and mixed with the bacteria. The resulting smear was covered with microscope slide. The analyzed images were taken with a Nikon 80i microscope.

Methods

Work-flow scheme

The work-flow scheme applied in this work consists of the following steps:

1. Segmentation of the Region of Interest.
2. Feature Calculation.
3. Feature Selection.
4. Class Recognition.

Segmentation of the Region of Interest

The aim of image segmentation is to separate the Region of Interest (ROI) from the background by creating a binary mask. In our case, ROI is the area where bacteria are located and this subarea of the mask is set to be white, while the background remains black. At first the image is converted to grayscale, then the Otsu method [18] combined with open and close morphological operations [20] is applied. These computations are performed with MATLAB functions: *rgb2gray*, *multithresh*, *imbinarize*, *imfill* and *bwareaopen*.

Calculation of geometric features

The shape of bacteria depends e.g. on their genera and growth phase. The geometric features are measured here on typical bacteria instances selected from each microscopic image and applied later for the classification purposes.

Dependencies between vectors. Let $\mathcal{Q}_m = \{q_k\}_{k=0}^m$ be a set of $m + 1$ planar points $q_k = (x_k, y_k)$ on a single bacteria's boundary in 2D-Euclidean space. These points are set in a clockwise order according to the following procedure. Recall that in MATLAB function $atan2(\tilde{y}, \tilde{x})$ calculates the angle between x-axis and a line joining point $\tilde{p} = (\tilde{x}, \tilde{y})$ with the origin of the coordinate system i.e. a point (0, 0). Upon shifting the origin to the point $c = (x_c, y_c)$ where $x_c = (1/(m+1)) \sum_{k=0}^m x_k$ and $y_c = (1/(m+1)) \sum_{k=0}^m y_k$ we applied here $atan2(x_k - x_c, y_k - y_c)$ —note that we also flipped variables in *atan2* to guarantee a clockwise order in $\mathcal{Q}_m$. The points are thus indexed in ascending order based on the *atan2* values. We pick now a point $q_{md} \in \mathcal{Q}_m$ whose Euclidean distance towards the point c is the smallest and then reorder points. If we have a sequence of elements $q_0, q_2, \ldots, q_m$ and we choose one of them as q_{md} it becomes the first element of the new sequence $\tilde{\mathcal{Q}}_m$. Then all elements following q_{md} are shifted after q_{md}, and finally

we append the elements that preceded q_{md} at the end of the sequence (so if we had q_0, q_1, q_2, q_3, q_4, q_5, q_6 and $q_{md} = q_3$ the new order reads as q_3, q_4, q_5, q_6, q_0, q_1, q_2). Next the set $\tilde{\mathcal{Q}}_m$ is reduced to $\hat{\mathcal{Q}}_n = \{\hat{q}_i\}_{i=0}^n$ upon picking $n + 1$ points. In this work $n + 1 = 10$ is arbitrarily selected for all bacteria. The points forming $\hat{\mathcal{Q}}_n$ are selected applying the following formula *fix*(*linspace*(0, *m*, *n* + 2)). Employing these functions provides a guarantee that the distances between the selected points are equal in terms of their indices, while minimizing the differences between these distances. Assume m is equal to 108 and $n + 2$ to 11, applying the *linspace* function results in the following values: 0, 10.8, 21.6, 32.4, 43.2, 54, 64.8, 75.6, 86.4, 97.2, 108. After processing by the *fix* function and omitting the first element, we obtain the indices of the points in $\tilde{\mathcal{Q}}_m$—10, 21, 32, 43, 54, 64, 75, 86, 97, 108, which form the set of points $\hat{\mathcal{Q}}_n$. In the next step we calculate distances between each $\hat{q}_i$ and $\hat{q}_{i+1}$ (and the distance between $\hat{q}_n$ and $\hat{q}_0$), and between each $\hat{q}_i$ and c.

The latter approach is illustrated in Fig 2. Note that no matter how the figure is rotated we always pick q_{md} placed in the corresponding similar position on bacteria's boundary resulting in a similar order of vector elements (starting with its q_{md}).

In addition, for a selected k-th bacteria based on its boundary points $\hat{\mathcal{Q}}_n^{(k)}$, a set of $n + 1$ triangles $\{\Delta_i^{(k)}\}_{i=0}^n$ is formed each determined by the vertices $\{\hat{q}_i^{(k)}, \hat{q}_{i+1}^{(k)}, c^{(k)}\}$ (the last triangle $\Delta_n^{(k)}$ is defined by $\{\hat{q}_n^{(k)}, \hat{q}_0^{(k)}, c^{(k)}\}$)—see Fig 2. Recalling that $\rho(x, y) = \sqrt{\sum_{i=0}^n (x_i - y_i)^2}$ defines the Euclidean distance for $x, y \in \mathbb{E}^n$, define now for each $\hat{\mathcal{Q}}_n^{(k)}$ (with $\vec{v}_{k,i} = \hat{q}_i^{(k)} - c^{(k)}$, $\vec{r}_{k,i} = c^{(k)} - \hat{q}_i^{(k)}$ and $\vec{w}_{k,i} = \hat{q}_i^{(k)} - \hat{q}_{i+1}^{(k)}$, for $i = 0, \ldots, n$; note $\vec{w}_{k,n} = \hat{q}_0^{(k)} - \hat{q}_n^{(k)}$) the following measures (see Fig 2):

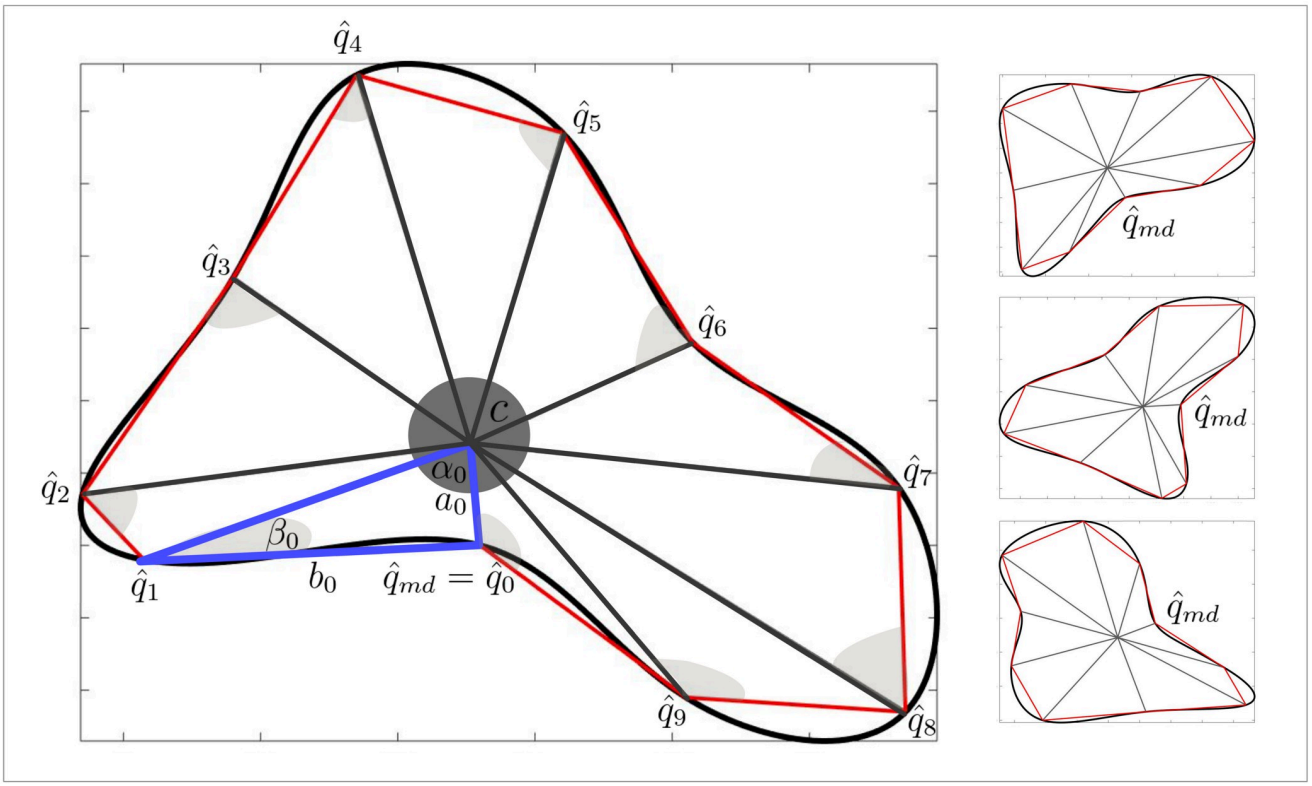

Fig 2. Distances and angles on exemplary shape, and exemplary rotations of a shape with applied method. Vectors: $\vec{a}_k$—dark gray lines, $\vec{b}_k$—red lines, $\vec{\alpha}_k$—dark gray angles, $\vec{\beta}_k$—light gray angles.

https://doi.org/10.1371/journal.pone.0293362.g002

- $\vec{a}_k$—vector of $n + 1$ distances $\rho(\hat{q}_i^{(k)}, c^{(k)})$,
- $\vec{b}_k$—vector of $n + 1$ distances $\rho(\hat{q}_i^{(k)}, \hat{q}_{i+1}^{(k)})$ (with $b_{k,n} = \rho(\hat{q}_n^{(k)}, \hat{q}_0^{(k)})$),
- $\vec{\alpha}_k$—vector of of $n + 1$ angles $\arccos(\vec{v}_{k,i+1}, \vec{v}_{k,i})$ (with $\alpha_{k,n} = \arccos(\vec{v}_{k,0}, \vec{v}_{k,n})$),
- $\vec{\beta}_k$—vector of $n + 1$ angles $\arccos(\vec{r}_{k,i+1}, \vec{w}_{k,i})$ (with $\beta_{k,n} = \arccos(\vec{r}_{k,0}, \vec{w}_{k,n})$).

In one microscopic image there might be hundreds of bacteria instances. Some of them are grouped together with overlaps which results in being identified as a single object once Region of Interest mask is applied. Another impeding factor comes from the fact that bacteria image can be taken at various stage of growth potentially related to its varying shape. Burying in mind the above concerns, only several b bacteria instances from the ROI mask are considered. These bacteria are selected based on the area value of the objects. All objects are sorted in ascending order and b items are selected with the area value closest to the median of all area values of the objects in a single image. Such approach ensures selection of objects representing single bacteria cells rather than groups of overlapped cells. In this research, for the calculation of features {1} and {2} we set $b = 50$ and for {3}, {4}, {5} and {6} $b = 10$ (for enumeration of features see subsection—All geometric features).

Each of the selected b bacteria on a given image represented by vector measures $(F_b = \{F_k\}_{k=1}^{b}$ where $F_k = (\vec{a}_k, \vec{b}_k, \vec{\alpha}_k, \vec{\beta}_k) \in \mathbb{R}^{4(n+1)})$ is compared with the exemplary bacteria measure selected by experts which is represented by $F_e = (\vec{a}_e, \vec{b}_e, \vec{\alpha}_e, \vec{\beta}_e) \in \mathbb{R}^{4(n+1)}$.

To illustrate the vector comparison procedure and to prove its credibility on more distinctive shapes an example of shape comparison between $F_{bs} = (\vec{a}_{bs}, \vec{b}_{bs}, \vec{\alpha}_{bs}, \vec{\beta}_{bs}) \in \mathbb{R}^{4(n+1)}$ with other vectors is presented (see Fig 3). F_{bs} is a set of vectors of values calculated for bacteria-shaped object. This object is an irregular oval shape that represents a bacteria cell. F_{bs} is compared with: F_{b2}—vectors calculated for bacteria-shaped object with double magnified size, F_h—vectors calculated for horseshoe shape, F_r—vectors calculated for a round shape and lastly, F_o—vectors calculated for oval shape. All shapes in question are artificially created with a slight irregularity applied. The latter corresponds to the objects selected by the ROI mask as they are also irregular and not the symmetric round or oval shapes.

For a given bacteria-like shape represented by F_{bs} we calculate the Pearson coefficient value [21] for all corresponding pairs of vectors in (F_{bs}, F_n), where $F_n \in \{F_{bs}, F_{b2}, F_h, F_r, F_o\}$, to verify how its value corresponds to the object shape Fig 3. Table 1 shows the correlation values between bacteria-shaped object and the same object resized, whereas Table 2 reports on correlation between bacteria-shaped object and other shapes. These calculations are conducted on 200×200 pixel images rotated by angles 0˚, 45˚, 90˚ and 135˚ in a counterclockwise direction around the center of the image. In this experiment, all F_n are calculated for each of the four selected angles yielding: $F_{n^{0°}}, F_{n^{45°}}, F_{n^{90°}}, F_{n^{135°}}$ (e.g. for round shape we have $F_{r^{0°}}, F_{r^{45°}}, F_{r^{90°}}, F_{r^{135°}}$). Note that F_n is equal to $F_{n^{0°}}$. Calculated data show significant impact of the object shape on the coefficient value for vectors of a and α, and that coefficient value is almost independent from the object size and rotation.

Despite the fact that Pearson coefficient properly reflects the relationship between shapes (expressed in vector forms and compared respectively), in the case of bacteria comparison, better classification results can be obtained upon replacing this coefficient with a slightly different approach presented in the following example.

To compare two vectors (e.g. representing some abstract feature), assume that the similarity between two vectors $w_1 = [1, 2, 3, 4]$ and $w_0 = [4, 1, 2, 3]$ is to be established. In doing so, the normalization of both vectors renders $w_{1n} = [0, 0.33, 0.66, 1]$ and $w_{0n} = [1, 0, 0.33, 0.66]$. Then

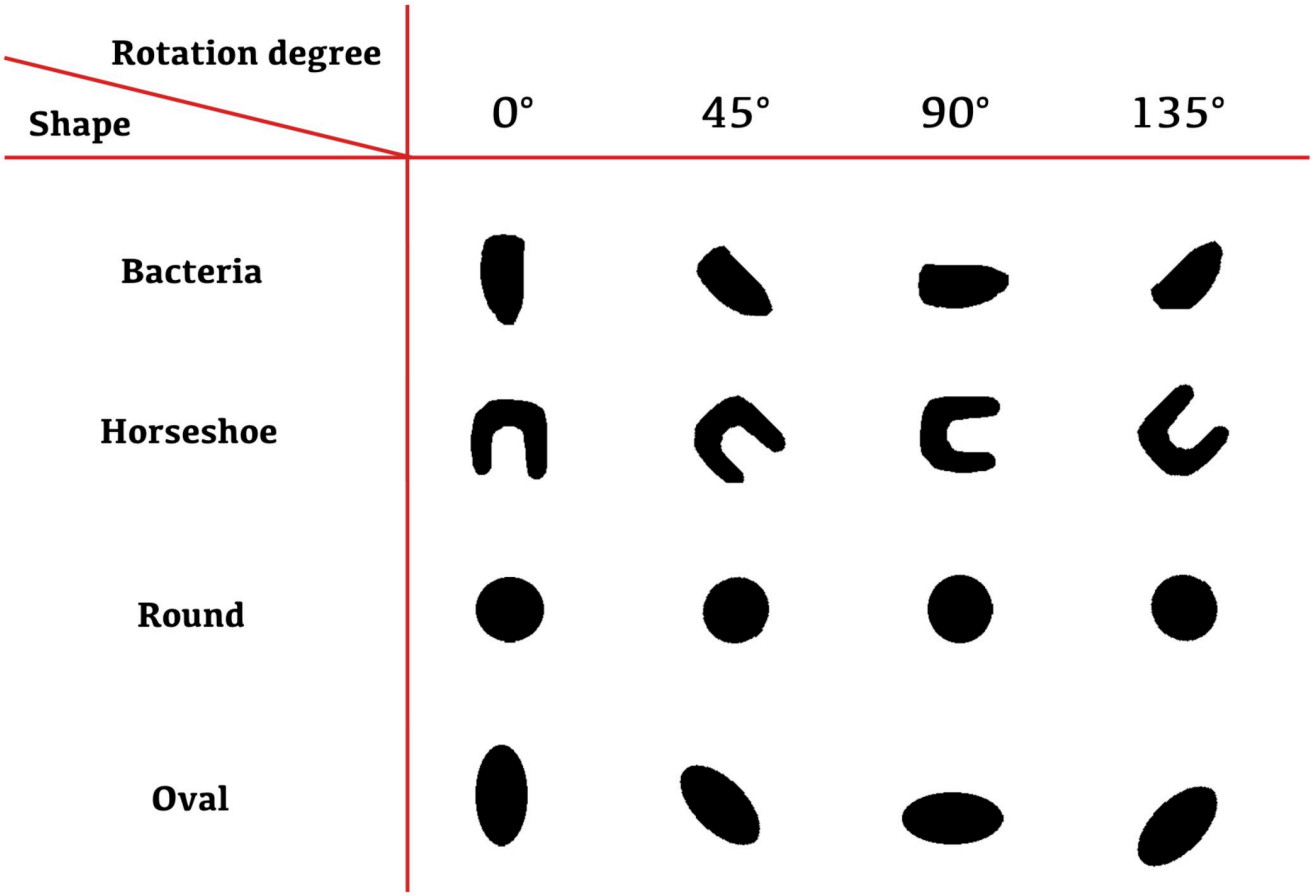

Fig 3. Shapes that were compared with a bacteria-like shape.

https://doi.org/10.1371/journal.pone.0293362.g003

three more vectors are created as they correspond to different positions of the object (on which w_1 was calculated): w_{2n} = [1, 0, 0.33, 0.66], w_{3n} = [0.66, 1, 0, 0.33] and w_{4n} = [0.33, 0.66, 1, 0]. In the next step, we calculate mean squared error [22] between each w_i (for $i = 1, \ldots, 4$) and w_0 which is equal to: 1.33, 0, 1.33 and 1.78, respectively. Then the smallest value is chosen, which here reads as 0 meaning that both w_{0n} and w_{in} are equal.

Vector w_1 can represent a certain vector calculated on currently analyzed shape (e.g. a_n from F_n) and w_0 stands for a corresponding vector calculated on a bacteria-like shape (e.g. a_{bs}

Table 1. Table presents Pearson coefficient between vectors calculated on bacteria-shaped object F_{bs} and the same object rotated ($F_{bs^{0°}}$, $F_{bs^{45°}}$, $F_{bs^{90°}}$, $F_{bs^{135°}}$) and F_{bs} with vectors calculated for bacteria-shaped object twice magnified rotated ($F_{b2^{0°}}$, $F_{b2^{45°}}$, $F_{b2^{90°}}$, $F_{b2^{135°}}$).

SV	$F_{bs^{0°}}$	$F_{bs^{45°}}$	$F_{bs^{90°}}$	$F_{bs^{135°}}$	$F_{b2^{0°}}$	$F_{b2^{45°}}$	$F_{b2^{90°}}$	$F_{b2^{135°}}$
$\vec{a}$	**1.00**	**0.98**	**1.00**	**0.98**	**0.99**	**0.98**	**0.99**	**0.98**
$\vec{b}$	1.00	0.30	1.00	0.30	0.03	0.36	0.03	0.32
$\vec{\alpha}$	**1.00**	**0.97**	**1.00**	**0.97**	**0.99**	**0.97**	**0.99**	**0.97**
$\vec{\beta}$	1.00	0.92	1.00	0.92	0.97	0.93	0.97	0.93

SV stands here for the Set of Vectors which is a set that consists of vectors corresponding to $\vec{a}$, $\vec{b}$, $\vec{\alpha}$ and $\vec{\beta}$ that are compared with the corresponding vectors in F_{bs} with the Pearson Coefficient.

https://doi.org/10.1371/journal.pone.0293362.t001

Table 2. Table presents pearson coefficient between vectors calculated on bacteria-shaped object F_{bs} with vectors calculated for horseshoe F_h, round F_r and oval F_o shapes rotated by 0˚, 45˚, 90˚ and 135˚.

SV	$F_{h^{0°}}$	$F_{h^{45°}}$	$F_{h^{90°}}$	$F_{h^{135°}}$	$F_{r^{0°}}$	$F_{r^{45°}}$	$F_{r^{90°}}$	$F_{r^{135°}}$	$F_{o^{0°}}$	$F_{o^{45°}}$	$F_{o^{90°}}$	$F_{o^{135°}}$
$\vec{a}$	**0.25**	**0.23**	**0.25**	**0.23**	**0.51**	**0.56**	**0.51**	**0.56**	**0.97**	**0.98**	**0.97**	**0.98**
$\vec{b}$	0.28	0.33	0.28	0.33	0.03	0.05	0.03	0.05	0.29	0.23	0.29	0.23
$\vec{\alpha}$	**0.07**	**0.16**	**0.07**	**0.16**	**0.46**	**0.60**	**0.46**	**0.60**	**0.91**	**0.92**	**0.91**	**0.92**
$\vec{\beta}$	0.19	0.19	0.19	0.19	0.20	0.35	0.20	0.35	0.88	0.90	0.88	0.90

SV stands here for the Set of Vectors which is a set that consists of vectors corresponding to $\vec{a}$, $\vec{b}$, $\vec{\alpha}$ and $\vec{\beta}$ that are compared with the corresponding vectors in F_{bs} with the Pearson Coefficient.

https://doi.org/10.1371/journal.pone.0293362.t002

from F_{bs}). Here all the shapes from Fig 3 are compared with the bacteria-like shape. Vectors w_1 and w_0 can also represent a vector calculated on currently analyzed bacteria (e.g. a_k from F_k) and on the exemplary one (e.g. a_e from F_e). These dependencies are calculated for every selected bacteria in the analyzed image.

Subsequently, the minimum mean square error value is computed for specific vectors corresponding to each of the b chosen bacteria in the image, across all four vectors. The corresponding results are denoted by $MSE\vec{a}_k^{min}, MSE\vec{b}_k^{min}, MSE\vec{\alpha}_k^{min}, MSE\vec{\beta}_k^{min}$, where $k = 1, 2, \ldots, b$. Then respective mean values of the minimum values for vectors $\vec{a}, \vec{b}, \vec{\alpha}, \vec{\beta}$ are calculated for all the selected bacteria from an analyzed image rendering four features based on geometry: $\overline{MSE\vec{a}^{min}}, \overline{MSE\vec{b}^{min}}, \overline{MSE\vec{\alpha}^{min}}, \overline{MSE\vec{\beta}^{min}}$.

Curvature and arc-length. Having selected $\hat{\mathcal{Q}}_n$ points (described in previous subsection) one can estimate the object's boundary with the aid of interpolation [23]. In order to define any interpolant γ which graph forms a closed curve the set $\hat{\mathcal{Q}}_n$ is augmented with an extra point $\hat{q}_{n+1} = \hat{q}_0$. The missing interpolation knots $\{\hat{t}_i\}_{i=0}^{n+1}$ for which $\hat{q}_i = \gamma(\hat{t}_i)$ are estimated from *exponential parameterization* [24, 25]:

$$\hat{t}_i = 0, \quad \hat{t}_{i+1} = \hat{t}_i + \|q_{i+1} - q_i\|^{\lambda}, \quad i = 0, \cdots, n$$

with $\lambda \in [0, 1]$. Here a special case of $\lambda = 0.5$ (the so-called *centripetal parameterization*) is used. Next a cubic spline $\gamma = \hat{\gamma}^{cs}$ with clamped boundary conditions [26] is applied (a complete spline). The latter requires an a priori information on $\hat{\gamma}'(\hat{t}_0) = v_0$ and $\hat{\gamma}'(\hat{t}_{n+1}) = v_{n+1}$ which is originally unavailable. In order to extract somehow v_0 and v_{n+1} an approach based on Modified Hermite scheme is used [27], where both v_0 and v_{n+1} are estimated from Lagrange Cubics $\hat{\gamma}_0^C$, $\hat{\gamma}_{n-2}^C$ fitting $\{\hat{q}_0, \hat{q}_1, \hat{q}_2, \hat{q}_3\}$ and $\{\hat{q}_{n-2}, \hat{q}_{n-1}, \hat{q}_n, \hat{q}_{n+1}\}$ yielding $v_0 = \hat{\gamma}_0^{C'}(\hat{t}_0)$ and $v_{n+1} = \hat{\gamma}_{n-2}^{C'}(\hat{t}_{n+1})$, respectively.

Having constructed a complete spline $\gamma = \hat{\gamma}^{cs}$ one can compute its curvature:

$$\kappa(t) = \frac{\|\vec{T}'(t)\|}{\|\vec{r}\,'(t)\|},$$

where $\vec{r}(t) = \dot{\gamma}(t)$ is a tangent vector to γ at t with its normalized vector $\vec{T}(t) = \vec{r}(t)/\|\vec{r}(t)\|$ or arc-length of the curve γ on interval $[a, t]$:

$$s = \int_a^t \|\vec{r}\,'(u)\|\, du.$$

All geometric features. The final set of features based on size and geometry of the selected objects reads as:

- {1} *Mean bacteria arc-length*—which is a sum of all $n + 1$ arc-lengths representing the perimeters of all selected bacteria divided by b,
- {2} *Mean curvature of b bacteria in one image*—a sum of all integrals of a curvature $\kappa(t)$ on each of the $[t_i, t_{i+1}] \ni t$ intervals calculated for each bacteria where $i = 0, 1, \ldots, n$. Then the sum of integrals is divided by b,
- {3} *Minimal mean square error first distance*—$\overline{MSE\vec{a}^{min}} = (1/b)\sum_{k=1}^{b} MSE\vec{a}_k^{min}$,
- {4} *Minimal mean square error second distance*—$\overline{MSE\vec{b}^{min}} = (1/b)\sum_{k=1}^{b} MSE\vec{b}_k^{min}$,
- {5} *Minimal mean square error first angle*—$\overline{MSE\vec{\alpha}^{min}} = (1/b)\sum_{k=1}^{b} MSE\vec{\alpha}_k^{min}$,
- {6} *Minimal mean square error second angle*—$\overline{MSE\vec{\beta}^{min}} = (1/b)\sum_{k=1}^{b} MSE\vec{\beta}_k^{min}$,
- {7} *Median of the object area in the image*,
- {8} *Percent of the bacteria area in the image*,
- {9} *Amount of objects in the image*—calculated sum of objects within the ROI mask,
- {10} *Amount of bacteria in the image*—calculated sum of the areas of objects within a ROI mask divided by the median of the object size in the current image.

Calculation of dispersion features

The dataset analyzed in this research consists of the images with bacteria monocultures. Each soil bacteria genera has a different colony dispersion. For some genera the bacteria cells are located close to each other in a non-uniform fashion, whereas the others are equally distributed. This section outlines the possible tools which measure the impact of such irregularities on classification in terms of mean shift [28], *k*-means [29] and regression [30].

Mean shift. Mean shift [28] is a scheme that allocates points through an iterative procedure to their average in a specified neighborhood (the local maxima of a density function) [31]. The output of this method consists of sets of points assigned to disjoint clusters determined by the distribution of input points. The resulting number of clusters in a clustering algorithm is determined by the algorithm itself. However, there are several input parameters that can be adjusted to customize the clustering process. These parameters include the window size, the distance metric used to evaluate the proximity of points to the cluster center and the stopping criteria for the algorithm. The mean shift algorithm flowchart is illustrated in Fig 4.

It is assumed here that the input data points of the mean shift algorithm are the centroids of the objects on the ROI mask captured with *props* MATLAB function. The generated features are the numbers of clusters to which the points were attached applying different values of r which is the radius of the window. The implementation of mean shift algorithm applied in this research can be found in MathWorks [32].

K-means. K-means [29] is a method that assigns points into k clusters. The algorithm is an iterative procedure of calculating distances between points and centroids, and shifting the centroids to new locations. The value of k is set arbitrarily. The flowchart of this algorithm is presented in Fig 5.

In order to determine features based on dispersion *k*-means method is firstly applied to cluster bacteria centroids. The latter incorporates their location in (x, y) coordinate system or both

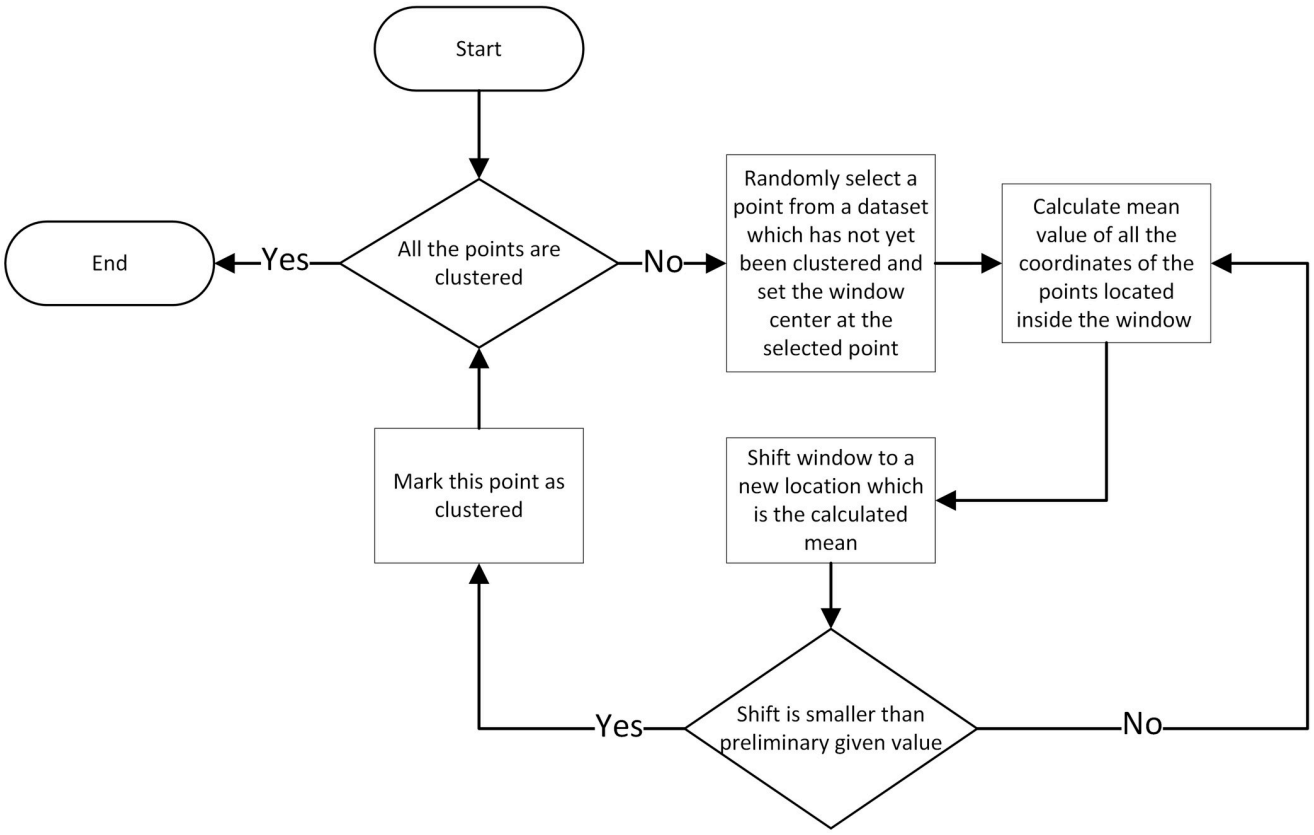

Fig 4. Flowchart of the mean shift algorithm.

https://doi.org/10.1371/journal.pone.0293362.g004

the Cartesian location combined with the area of the bacteria cell represented by (x, y, s). Assume the points $\mathcal{P}_{z_j} = \{p_i\}_{i=0}^{z_j}$ are given, where $z_j + 1$ defines the amount of points associated with the centroid c. Then a linear regression line is fitted to all points from $\mathcal{P}_{z_j}$. Let $\mathcal{Q}_{z_j} = \{q_i\}_{i=0}^{z_j}$ be the points on the fitted linear regression line such that $q_i = (x_{q_i}, y_{q_i})$ and each $x_{q_i} = x_i$

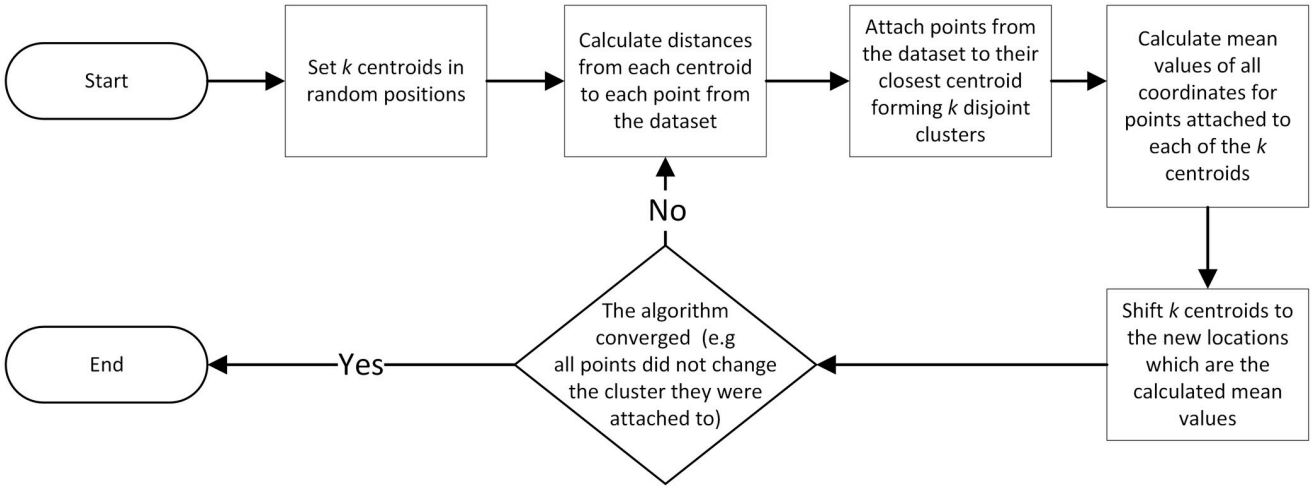

Fig 5. Flowchart of the K-means algorithm.

https://doi.org/10.1371/journal.pone.0293362.g005

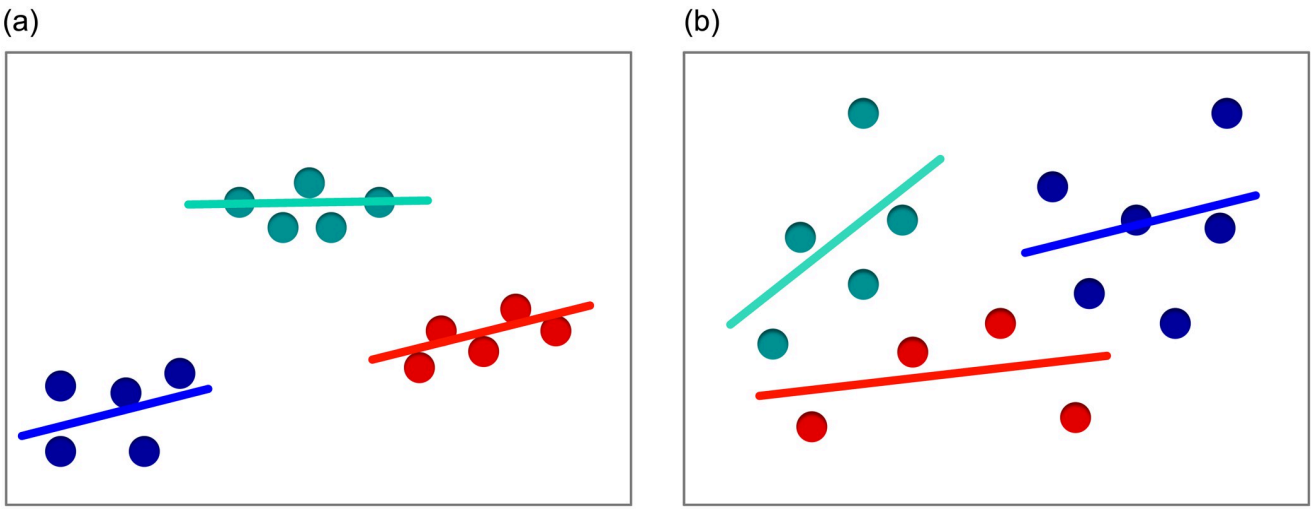

Fig 6. K-means algorithm and linear regression image with m + 1 = 15 for k = 3 put in sets (a) and evenly spaced (b).

https://doi.org/10.1371/journal.pone.0293362.g006

for x_i being first coordinate of the point $p_i \in \mathcal{P}_{z_j}$. Next $\bar{d}_j = (1/(z_j + 1)) \sum_{i=0}^{z_j} \|y_i - y_{q_i}\|$, which is the mean distance between each of the corresponding points p_i and q_i for j'th centroid, is calculated. Note that $\bar{d}_j$ can also be weighted by the values of the normalized vector of bacteria surface areas s_i computed as $\bar{d}_j = (1/\sum_{i=0}^{z_j} s_i) \sum_{i=0}^{z_j} \|y_i - y_{q_i}\| s_i$. Such procedure is repeated for each of the k clusters. The resulting sum $\bar{D} = \sum_{j=1}^{k} \bar{d}_j$ becomes the feature value for currently analyzed image.

To provide a clear example, consider two sets of $m + 1 = 15$ points. The first set is composed of points grouped into three subsets, while the second set contains evenly spaced samples. The points in both sets are attached to $k = 3$ clusters by k-means algorithm. Next, one calculates the values of $\bar{D}$ for both datasets as specified in the preceding paragraph. The computed values of $\bar{D}$ for images from Fig 6 are equal to $\bar{D} = 68$ for Fig 6a and $\bar{D} = 1058$ for Fig 6b. A marked discrepancy is observed in the results depending on the level of data dispersion. Fig 7 illustrates the latter used for the exemplary microscopic images.

All dispersion features. The final set of features based on the size and geometry consists of:

- {11-18} *Mean shift*—for different r values equal to = 25, 50, 75, 100, 125, 150, 175, 200, respectively,
- {19-27} *K-means and regression*—for (x, y) with $k = 2, 6, 10$, for (x, y, s) with $k = 2, 6, 10$ and for (x, y) with $k = 2, 6, 10$ weighted.

Calculation of luminance and chrominance features

In this work, the classification results obtained applying features based on geometry and dispersion is compared with an accuracy rendered by features based on chrominance and luminance. In doing so, statistical measures of the pixel values, i.e. colors defined by RGB (red, green and blue) color space on the whole image or only on the area covered with the ROI mask are computed. Such features are calculated either on the image converted to grayscale or within a selected color channel. The four statistical measures employed here to analyze the

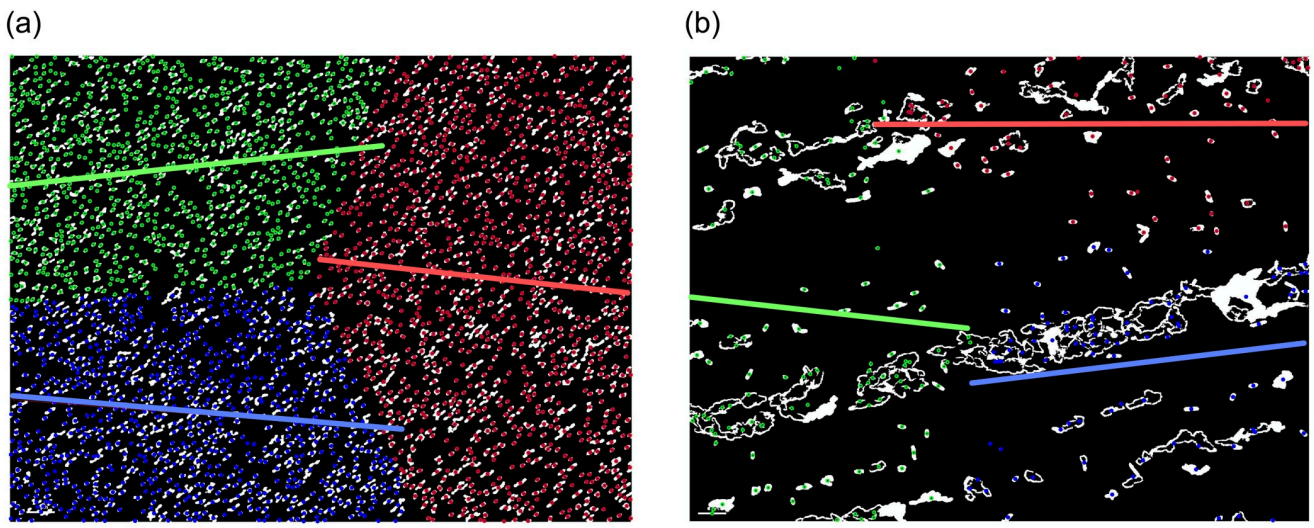

Fig 7. K-means algorithm and linear regression for microscopic image of *Rhizobium* (a) and *Enterobacter* (b).

https://doi.org/10.1371/journal.pone.0293362.g007

data are: variance [33], mean, kurtosis and skewness [34]. The resulting set of features based on color consists of:

- {28-35} *Variance*—calculated on mask {28}, on whole image {29}, on whole image red {30}, green {31}, blue channel {32}, calculated on mask for red {33}, green {34}, blue channel {35},
- {36-43} *Mean*—calculated on mask {36}, on whole image {37}, on whole image red {38}, green {39}, blue channel {40}, calculated on mask for red {41}, green {42}, blue channel {43},
- {44-51} *Kurtosis*—calculated on whole image {44}, on the mask {45}, on whole image red {46}, green {47}, blue channel {48}, on the mask for red {49}, green {50}, blue channel {51},
- {52-59} *Skewness*—calculated on whole image {52}, on the mask {53}, on whole image red {54}, green {55}, blue channel {56}, on the mask for red {57}, green {58}, blue channel {59}.

Feature selection

Noticeably, not all calculated features are appropriate for the classification. Some of them are not highly correlated with the affiliation to the class or their correlation with other features is too high which might cause redundancy. Such features should not be considered in the stage of class recognition. The feature selection methods solve this problem by picking appropriate features. In this work, we decided to verify the results given by the following methods:

- Fast Correlation Based Filter (FCBF) [35],
- Sparse Multinomial Logistic Regression with Bayesian Regularization (SBMLR) [36],
- Correlation-based Feature Selection (CFS) [37].

Class recognition

Class recognition methods are used here to assign input images to certain classes representing different bacteria genera. These methods are trained on the training set and their classification

performance is measured upon applying the testing set. Such sets contain selected features calculated for each of the images. Class recognition methods considered here include:

- Support Vector Machine (SVM) [38],
- Random Forest (RF) [39],
- K-Nearest Neighbors (KNN) [40],
- Multi-Layer Perceptron (MLP) [41].

These methods representing classical machine learning techniques rely on admitting features a priori determined by hand. Such class recognition methods continue to be widely used across a diverse range of applications [42]. In particular, these AI tools are also studied in the context of soil microorganism classification with high accuracy results reported [18].

Random forest. Random forest [39] is a group learning method whose task is to generate a set of models—trees, and then to classify the tested object into one of the classes taking into consideration the results from individual models. The trees are built based on the features table with known class assignment (supervised learning). Each node of the tree has conditions for numeric or non-numeric data. Satisfying these conditions determines object affiliation to one of the classes by the current model. In order to create a decision tree for RF (based on a table of features) one firstly randomly selects a subset of samples (table rows) with repetitions and places them into a so-called bootstrap dataset (it has as many rows as the input table of features) [43]. Having created the new dataset we draw from it x features (table columns) and verify which one will be the best for building the model (correctly separates the samples). The decision on which of the x features is to be used at a given tree node is made on the basis of methods such as e.g. Gini Impurity or Entropy [43]. The same measures allow us to set a threshold for condition concerning numeric data for a given feature. For the classification purposes hundreds of trees are generated. Upon creation of n trees one verifies to which class a new instance is assigned by each of the models. The final decision on the classification is made according to the majority voting rule. The effectiveness of this method is examined by comparing the achieved affiliation to a class by means of the algorithm with the actual instance assignment. One can arbitrarily select the value of n, however, with n increasing, the computational complexity of the algorithm explodes, resulting in a longer computation time. In this work the *TreeBagger* MATLAB function was applied.

Results

The dataset considered here [44] consists of 128 microscopic images of soil bacteria from the five selected genera: *Enterobacter*—22 images, *Rhizobium*—25 images, *Pantoea*—26 images, *Bradyrhizobium*—34 images and *Pseudomonas*—21 images. These images have not been preprocessed either by the microbiologists (no chemical reactions conducted) or by any computerized system. In the experimental section the concept of cross validation [45] is applied. More specifically, 10% ratio cross validation is used, in which the set of images is randomly shuffled and divided into ten subsets. Next, nine of these sets are selected to form the training set on which our model learns how to distinguish input objects among different classes. The remaining set (called the testing set) is used to verify how good the result of classification is by calculating its accuracy. The model accuracy represents the amount of correctly classified bacteria images divided by the amount of the whole set of images (in the testing set). Then another of the ten sets becomes the testing set so that we have ten iterations (ten different training and testing sets) and calculate the mean accuracy value of ten iterations. The tables in this section display the mean accuracy resulting from 50 iterations of 10% cross-validation.

Table 3. The accuracy obtained with different feature selection and classification methods performed on features based on geometry and dispersion for the five bacteria genera.

FSM	SVM	RF (n = 200)	KNN (k = 1)	MLP	Selected features
None	78.3438	**85.1406**	80.5938	50.9375	{1-27}
FCBF	75.0312	82.2656	79.0469	63.5	{7, 11, 2, 5, 8}
SBMLR	40.8906	36.2344	36.2344	28.1094	{1}
CFS	78.3906	83.875	79.8594	60.1719	{1, 2, 5-7, 10, 11, 14}

FSM stands here for the Feature Selection Method.

https://doi.org/10.1371/journal.pone.0293362.t003

The final results include calculations based on iteratively selected parameters of class recognition methods which are: Support Vector Machine (with default parameters in *fitcsvm* MATLAB function), Random Forest with 200 trees, K-Nearest Neighbors with $k = 1$ and Multi-Layer Perceptron with network topology 15 – 15 – 15 trained with backpropagation algorithm based on gradient descent. The parameters were selected to maximize the ACC.

The accuracy for the whole set of features consisting of geometry and dispersion traits shown in Table 3 reached 85.14% for Random Forest for the whole set of five different bacteria genera. Applying feature selection methods does not increase the achieved result. The results for four different bacteria genera presented in Table 4 are also the highest for Random Forest ranging from 81.7% to 91.6%.

Table 5 presents accuracy for different sets and subsets of features. Features based on dispersion obtained the best accuracy of 63.72% for KNN, whereas features based on geometry reached 82.59% for Random Forest. Combining these sets increases the result by 2.55

Table 4. The accuracy computed with different classification methods performed on the whole set of features based on geometry and dispersion for four selected bacteria genera (subsets of five bacteria genera).

Selected bacteria genera	SVM	RF (n = 200)	KNN (k = 1)	MLP
1, 2, 3, 4	81.7944	86.2243	85.9252	58.8785
2, 3, 4, 5	76.1132	81.6792	80.3208	54.9057
3, 4, 5, 1	75.8641	84.1748	78.6408	51.9223
4, 5, 1, 2	88.3922	88.4314	86.8235	64.8627
5, 1, 2, 3	86.0638	**91.5532**	87.8298	69.617

1, 2, 3, 4 and 5 stand for the following bacteria genera respectively: *Enterobacter*, *Rhizobium*, *Pantoea*, *Bradyrhizobium* and *Pseudomonas*.

https://doi.org/10.1371/journal.pone.0293362.t004

Table 5. The accuracy obtained with different classification methods performed on different sets of features based on color, geometry and dispersion (and their subsets).

Set of features based on. . .	SVM	RF (n = 200)	KNN (k = 1)	MLP
geometry and dispersion {1-27}	78.3438	**85.1406**	80.5938	50.9375
geometry {1-10}	74.7969	**82.5938**	70.9062	53.0312
vectors (from geometry) {3-6}	47.5469	**48.5469**	41.9844	32.0625
dispersion {11-27}	60.4531	63.4219	**63.7188**	42.5312
k-means (from dispersion) {19-27}	39.8281	**46.8594**	43.7188	24.7976
mean shift (from dispersion) {11-18}	53.7188	**63.9375**	63.2812	49.9219
color {28-59}	86.2188	94.2969	**95.5938**	88.3906
color, geometry and dispersion {1-59}	89.7969	**94.8281**	91.2031	75.4688

https://doi.org/10.1371/journal.pone.0293362.t005

Table 6. The accuracy with different classification methods for the five bacteria genera performed on features based on k-means and regression. The set consists of 60 features, for k = 1, . . ., 20 and 2 dimensional vector for k-means, 2 dimensional vector for k-means with weighted variance and 3 dimensional vector for k-means.

Quantile	SVM	RF (n = 200)	KNN (k = 1)	MLP
0	48.0469	58.9062	48.2031	28.2031
0.2	45.7812	60.625	41.875	26.5625
0.4	50.8594	**61.7969**	44.5312	26.0938
0.6	40	50.3125	36.6406	23.5938

The value in Quantile column informs that the bacteria were taken into account if their area value was greater then certain quantile of area value of the bacteria on a given image.

https://doi.org/10.1371/journal.pone.0293362.t006

percentage points and amounts to 85.14% for Random Forest. One can also analyze the results of the selected subsets of features based on dispersion and geometry. As an example, features extracted based on mean shift yields up to 63.94% accuracy, whereas features based on *k*-means render 46.86%. Applying features based on vectors reached only 46.86% accuracy; however, one of them—feature number 5—is accepted by both FCBF and CFS feature selection methods which is shown in Table 3 what proves its significant impact on increasing the classification accuracy.

The features calculated with *k*-means may seem insignificant. For this reason the classification results are presented depending on the amount of bacteria analyzed on a single image based on their area value. For each of the calculations, as shown in Table 6, a different quantile value is selected which means that analyzed bacteria are ones which area exceeds or is equal to that quantile area value in a single image. The 60 features were calculated for *k*-means for each of the quantiles: 0, 0.2, 0.4 and 0.6. For example for quantile equal to 0.2 the calculated features are: for k = 1, . . ., 20 and 2 dimensional vector for k-means, 2 dimensional vector for k-means with weighted variance and 3 dimensional vector for k-means—yielding 60 features for this quantile. The highest results are reached for the quantile equal to 0.4 amounting to 61.8%. It is remarked here that extending the final set of features by these 60 *k*-means features does not improve the final result. For that sheer reason only nine previously calculated features based on *k*-means are chosen.

In this research the highest classification accuracy for a set of geometry and dispersion features yields 85.14%. In previous work [19], based on the same image data set, the accuracy obtained amounts to 97%. The latter analyses different set of features, involving geometric and texture characteristics. The texture features rely on luminance and chrominance, which may artificially improve the accuracy of the results. For example, this may occur when microscopic images from each genera are taken under different lighting conditions. Thus, the obtained accuracy 85.14% forms a promising result as the examined features are not based on color information.

Discussion

The classification based on extracting features from bacteria geometry and dispersion yields a promising 85.14% ACC. The latter is reached for the Random Forest classifier to identify five selected soil bacteria genera. The experiments conducted on features based on geometry and dispersion separately rendered 82.59% in case of Random Forest and 63.72% for K-Nearest Neighbors. These results illustrate that applying a proper set of features with no color traits enables classification of soil bacteria. The latter permits to bypass the impact of lighting conditions and coloring of samples on classification. In contrast the geometry and dispersion based

classification is insensitive to the last two factors. However, the difference between the classification accuracy based on geometry and dispersion traits versus this one based on color traits is significant (around 10 percentage points) and there are some issues requiring future research investigation.

Indeed, one needs to apply a different method of selecting points on bacteria boundary to highlight the characteristic elements of its shape. In addition, various parameterizations to estimate the unknown interpolation knots combined with different interpolation fitting schemes might also be considered [46]. Other methods for object dispersion in the image should also be examined. In this work we compared the results given by the four classification methods: Support Vector Machine, Random Forest, K-Nearest Neighbors and Multi-Layer Perceptron. Other classifiers such as Extreme Learning Machines or Deep Learning Methods may provide more effective recognition tools. The features in the future research can be also computed applying Convolutional Neural Networks [47].

The generated results are calculated on the dataset with a single bacteria genera on an input image. These organisms were grown under laboratory conditions, with no contamination involved (as they are all immersed in uniform medium). In future research, the testing should also be performed on images taken from the natural environment (e.g. from the genuine rizosphere sample). The ultimate goal is to classify different bacteria genera mixed with extra organic or non-organic objects as they cohabit in a real soil sample. More importantly, the classification results on images that contain different bacteria genera (for example mixes of two or three genera on one image) should also be examined. In particular, the final recognition tool should allow to assess the quantity of bacteria cells affiliated to a certain genera on the currently analyzed microscopic image.

The classification system created in this work can be applied in practice. However, further research is needed for samples containing strains of different species of bacteria representing the same genus. These species differ in phenotypic features (morphological, biochemical and physiological). The number of analyzed strains of bacteria has an important meaning. We are unable to draw a conclusion from a single photograph of cells or bacterial colonies known to be of some type of bacteria. As an example, the genus *Pseudomonas* includes both fluorescent and non-fluorescent bacterial species. Problems with identifying bacteria based on their morphology result from reasons such as: (i) the influence of the environment, i.e. the composition of the medium and incubation time on the cell morphology, (ii) the phase of the bacterial cell cycle, (iii) the common morphology of cells of different types of bacteria. It is worth mentioning that so far there are over 10 thousand species of culturable bacteria, with a huge number of species that cannot be cultured in the laboratory. It is very important to accurately classify the bacteria as a representative of the appropriate species. The latter permits to decide whether to use it for utilitarian purpose e.g. in biological protection of plants against diseases or to apply suitable control against a given organism if it causes diseases (pathogen) or is harmful in any other respect.

Author Contributions

Conceptualization: Aleksandra Konopka, Ryszard Kozera.

Data curation: Aleksandra Konopka.

Formal analysis: Aleksandra Konopka.

Methodology: Aleksandra Konopka.

Resources: Lidia Sas-Paszt, Pawel Trzcinski, Anna Lisek.

Software: Aleksandra Konopka.

Visualization: Aleksandra Konopka.

Writing – original draft: Aleksandra Konopka, Ryszard Kozera.

Writing – review & editing: Aleksandra Konopka, Ryszard Kozera.

References

1. Sharma A, Lee S, Park YS. Molecular typing tools for identifying and characterizing lactic acid bacteria: a review. Food Sci Biotechnol. 2020; 29(10):1301–1318. https://doi.org/10.1007/s10068-020-00802-x PMID: 32995049
2. Maiden MC, Van Rensburg MJJ, Bray JE, Earle SG, Ford SA, Jolley KA, et al. MLST revisited: the gene-by-gene approach to bacterial genomics. Nat Rev Microbiol. 2013; 11(10):728–736. https://doi.org/10.1038/nrmicro3093 PMID: 23979428
3. Numberger D, Ganzert L, Zoccarato L, Mühldorfer K, Sauer S, Grossart HP, et al. Characterization of bacterial communities in wastewater with enhanced taxonomic resolution by full-length 16S rRNA sequencing. Sci Rep. 2019; 9. https://doi.org/10.1038/s41598-019-46015-z PMID: 31273307
4. Church DL, Cerutti L, Gürtler A, Griener T, Zelazny A, Emler S. Performance and application of 16S rRNA gene cycle sequencing for routine identification of bacteria in the clinical microbiology laboratory. Clin Microbiol Rev. 2020; 33(4). https://doi.org/10.1128/CMR.00053-19 PMID: 32907806
5. Floridia-Yapur N, Rusman F, Diosque P, Tomasini N. Genome data vs MLST for exploring intraspecific evolutionary history in bacteria: much is not always better. Infect Genet Evol. 2021; 93. https://doi.org/10.1016/j.meegid.2021.104990 PMID: 34224899
6. Caprette DR. Describing Colony Morphology [Internet]; 2022 [cited 2022 Sep 13]. Available from: https://bit.ly/324cqkA.
7. Tarca AL, Carey VJ, Chen X, Romero R, Drăghici S. Machine learning and its applications to biology. PLoS Comput Biol. 2007; 3(6):e116. https://doi.org/10.1371/journal.pcbi.0030116 PMID: 17604446
8. Rani P, Kotwal S, Manhas J, Sharma V, Sharma S. Machine learning and deep learning based computational approaches in automatic microorganisms image recognition: methodologies, challenges, and developments. Arch Comput Methods Eng. 2021. https://doi.org/10.1007/s11831-021-09639-x PMID: 34483651
9. Khutlang R, Krishnan S, Dendere R, Whitelaw A, Veropoulos K, Learmonth G, et al. Classification of mycobacterium tuberculosis in images of ZN-stained sputum smears. IEEE trans inf technol. 2010; 14 (4):949–957. https://doi.org/10.1109/TITB.2009.2028339 PMID: 19726269
10. Kang R, Park B, Eady M, Ouyang Q, Chen K. Single-cell classification of foodborne pathogens using hyperspectral microscope imaging coupled with deep learning frameworks. Sensors and Actuators B: Chemical. 2020; 309:127789. https://doi.org/10.1016/j.snb.2020.127789
11. Hiremath PS, Bannigidad P. Identification and classification of cocci bacterial cells in digital microscopic images. Int J Comput Biol. 2011; 4(3):262. https://doi.org/10.1504/IJCBDD.2011.041414 PMID: 21778559
12. Kotwal S, Rani P, Arif T, Manhas J, Sharma S. Automated bacterial classifications using machine learning based computational techniques: architectures, challenges and open research issues. Arch Computat Methods Eng. 2022; 29:2469–2490. https://doi.org/10.1007/s11831-021-09660-0 PMID: 34658617
13. Liu J, Dazzo F, Glagoleva O, Yu B, Jain A. CMEIAS: a computer-aided system for the image analysis of bacterial morphotypes in microbial communities. Microb Ecol. 2001. https://doi.org/10.1007/s002480000004 PMID: 11391457
14. Ruusuvuori P, Seppälä J, Erkkilä T, Lehmussola A, Puhakka JA, Yli-Harja OP. Efficient automated method for image-based classification of microbial cells. 19th Int Conf Pattern Recognit. 2008.
15. Ducret A, Quardokus E, Brun Y. MicrobeJ, a tool for high throughput bacterial cell detection and quantitative analysis. Nat Microbiol. 2016. https://doi.org/10.1038/nmicrobiol.2016.77 PMID: 27572972
16. Batt CA, Tortorello ML. Encyclopedia of Food Microbiology. Academic Press; 2014.
17. Beeckmans S, Xie J. Glyoxylate cycle. Reference Module in Biomedical Sciences. 2015. https://doi.org/10.1016/B978-0-12-801238-3.02440-5

18. Kruk M, Kozera R, Osowski S, Trzciński P, Sas-Paszt L, Sumorok B, et al. Computerized classification system for the identification of soil microorganisms. Appl Math Inf. 2016; 10:21–31. https://doi.org/10.18576/amis/100103

19. Konopka A, Struniawski K, Kozera R, Trzciński P, Sas-Paszt L, Lisek A, et al. Classification of soil bacteria based on machine learning and image processing. In: ICCS 2022. Springer International Publishing; 2022. p. 263–277.

20. Soille P. Morphological image analysis: principles and applications. Springer-Verlag; 1999.

21. Okwonu FZ, Asaju BL, Irimisose AF. Breakdown analysis of pearson correlation coefficient and robust correlation methods. IOP Conference Series: Materials Science and Engineering. 2020;917(1).

22. Hodson TO, Over TM, Foks SS. Mean squared error, deconstructed. J Adv Model Earth Syst. 2021; 13 (12). https://doi.org/10.1029/2021MS002681

23. Caruso C, Quarta F. Interpolation methods comparison. Comput Math Appl. 1998; 35(12):109–126.

24. Kvasov B. Methods of shape-preserving spline approximation. World Scientific; 2000.

25. Kozera R, Noakes L, Wilkołazka M. Exponential parameterization to fit reduced data. Appl Math Comput. 2021; 391:1–19.

26. McClarren RG. Computational nuclear engineering and radiological science using python; 2018.

27. Kozera R. Curve modeling via interpolation based on multidimensional reduced data. Studia Informatica. 2004; 25(4B):1–140.

28. Fukunaga K, Hostetler L. The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Trans Inf Theory. 1975; 21(1):32–40. https://doi.org/10.1109/TIT.1975.1055330

29. Pérez-Ortega J, Almanza-Ortega NN, Vega-Villalobos A, Pazos-Rangel R, Zavala-Díaz C, Martínez-Rebollar A. The k-means algorithm evolution, in introduction to data science and machine learning. IntechOpen. 2019.

30. Schneider A, Hommel G, Blettner M. Linear regression analysis: part 14 of a series on evaluation of scientific publications. Dtsch Arztebl Int. 2010; 107:776–82. https://doi.org/10.3238/arztebl.2010.0776 PMID: 21116397

31. Cheng Y. Mean shift, mode seeking, and clustering. IEEE PAMI. 1995; 17(8):790–799. https://doi.org/10.1109/34.400568

32. Finkston B. Mean Shift Clustering, MATLAB Central File Exchange.; 2023 [cited 2023 Jan 4]. Available from: https://www.mathworks.com/matlabcentral/fileexchange/10161-mean-shift-clustering.

33. Wasserman L. All of statistics: a concise course in statistical inference. Springer texts in statistics; 2005.

34. Joanes DN, Gill CA. Comparing measures of sample skewness and kurtosis. J R Stat Soc. 1998; 47 (1):183–189.

35. Yu L, Liu H. Feature selection for high-dimensional data: a Fast Correlation-Based Filter solution. In: Proceedings, Twentieth International Conference on Machine Learning. vol. 2; 2003. p. 856–863.

36. Cawley G, Talbot N, Girolami M. Sparse Multinomial Logistic Regression via Bayesian L1 Regularisation. In: Advances in Neural Information Processing Systems. vol. 19. MIT Press; 2006. p. 209–216. Available from: https://proceedings.neurips.cc/paper/2006/file/b22b257ad0519d4500539da3c8bcf4dd-Paper.pdf.

37. Hall M. Correlation-based feature selection for machine learning [Ph.D. thesis]. The University od Waikato, Hamilton, New Zealand; 2000.

38. Cervantes J, Garcia-Lamont F, Rodríguez-Mazahua L, Lopez A. A comprehensive survey on support vector machine classification: applications, challenges and trends. Neurocomputing. 2020; 408:189–215. https://doi.org/10.1016/j.neucom.2019.10.118

39. Ho TK. Random decision forests. In: Proceedings of 3rd International Conference on Document Analysis and Recognition. vol. 1; 1995. p. 278–282.

40. Fix E, Hodges JL. Discriminatory analysis. Nonparametric discrimination: consistency properties. USAF School of Aviation Medicine. 1951;.

41. Popescu MC, Balas V, Perescu-Popescu L, Mastorakis N. Multilayer perceptron and neural networks. WSEAS Transactions on Circuits and Systems. 2009; 8.

42. Halim Z, Hussain S, Hashim Ali R. Identifying content unaware features influencing popularity of videos on YouTube: A study based on seven regions. Expert Syst Appl. 2022. https://doi.org/10.1016/j.eswa.2022.117836

43. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning: with applications in R. Springer; 2013.

44. Konopka A. Zenodo Repository: scientistit/PLOS-2023: v1.0.0; 2023 [cited 2023 Aug 18]. Available from: https://zenodo.org/record/7789436.

45. Stone M. Cross-validatory choice and assessment of statistical predictions. J R Stat Soc Series B Stat Methodol. 1974; 36(2):111–133.

46. de Boor C. A practical guide to splines. Springer; 2001.

47. Zieliński B, Plichta A, Misztal K, Spurek P, Brzychczy-Włoch M, Ochońska D. Deep learning approach to bacterial colony classification. PloS ONE. 2017. https://doi.org/10.1371/journal.pone.0184554 PMID: 28910352