Annotated dataset for deep-learning-based bacterial colony detection

László Makrai
Department of Microbiology and Infectious Diseases, University of Veterinary Medicine, 1143 Budapest, Hungary https://orcid.org/0000-0002-7273-9133

Bettina Fodróczy
Department of Microbiology and Infectious Diseases, University of Veterinary Medicine, 1143 Budapest, Hungary

Sára Ágnes Nagy
Centre for Bioinformatics, University of Veterinary Medicine, 1078 Budapest, Hungary
https://orcid.org/0000-0003-0411-5887

Péter Czeiszing
Department of Microbiology and Infectious Diseases, University of Veterinary M

István Csabai
Department of Physics of Complex Systems, Eötvös Loránd University, 1117 Budapest, Hungary
https://orcid.org/0000-0001-9232-9898

Géza Szita
Centre for Bioinformatics, University of Veterinary Medicine, 1078 Budapest, Hungary

Norbert Solymosi (solymosi.norbert@gmail.com)
Centre for Bioinformatics, University of Veterinary Medicine, 1078 Budapest, Hungary
https://orcid.org/0000-0003-1783-2041

Data Note

Keywords: bacteria, colony detection, deep learning, annotation,

Posted Date: May 2nd, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2873282/v1

Annotated dataset for deep-learning-based bacterial colony detection

László Makrai[1], Bettina Fodróczy[1,2], Sára Ágnes Nagy[2], Péter Czeiszing[1,2], István Csabai[3], Géza Szita[2], and Norbert Solymosi[2,3*]

[1]Department of Microbiology and Infectious Diseases, University of Veterinary Medicine, 1143 Budapest, Hungary
[2]Centre for Bioinformatics, University of Veterinary Medicine, 1078 Budapest, Hungary
[3]Department of Physics of Complex Systems, Eötvös Loránd University, 1117 Budapest, Hungary
[*]corresponding author(s): Norbert Solymosi (solymosi.norbert@gmail.com)

ABSTRACT

Quantifying bacteria per unit mass or volume is a common task in various fields of microbiology (e.g., infectiology, and food hygiene). Most bacteria can be grown on culture media. The unicellular bacteria reproduce by dividing into two cells, which increases the number of bacteria in the population. Methodologically, this can be followed by culture procedures, which mostly involve determining the number of bacterial colonies on the solid culture media that are visible to the naked eye. However, it is a time-consuming and laborious professional activity. Addressing the automation of colony counting by convolutional neural networks in our work, we have cultured 24 bacteria species of veterinary importance with different concentrations on solid media. A total of 57,028 colonies were annotated manually by bounding boxes on the 373 digital images of bacterial cultures. The published database will help developments that use artificial intelligence to automate the counting of bacterial colonies.

Background & Summary

In microbiology, the colony-forming unit (CFU) is used to determine the number of viable bacteria that can grow on solid media.[1] In all cases, CFU values can only be interpreted when normalized to a unit volume (e.g., ml). Knowledge of CFU is essential in clinical microbiology, food hygiene, and vaccine research. The CFU count is most commonly estimated by counting the number of colonies on solid culture media. As the determination of the number of living bacteria is often a key, but at the same time, the process of colony counting is rather time-consuming and labor-intensive, there are several attempts in the literature to automate the procedure. A number of tools (EBImage[2], ImageJ[3], OpenCFU[4], AutoCellSeg[5], CFUCounter[6]) have been developed and are used for colony counting, which has some predefined threshold (e.g., color) and counts the resulting objects. Although they can be of great help in laboratory work, it is important to be aware of their drawbacks. A general limitation of these solutions is that objects in the image that are not colonies (e.g., pieces of the wall of a Petri dish, air bubbles) may also appear in the result as colonies. Although some tools allow these erroneous detections to be corrected manually, this again requires time-consuming expert work. Also limiting their everyday use is that most of them cannot count colonies if the number of colonies in the Petri dish is too high.[6]

The use of artificial intelligence (AI) to automate colony counting seems obvious. In the AI approach, colony counting is first an object detection task. A further task could be the differentiation of bacterial species, which requires a classification solution. We can obtain the total and per-class CFU count by counting the detected and classified objects. The number of colonies detected and classified can then be used to estimate the total, and per-species CFU counts. There are several machine-learning approaches available to solve this kind of problem. Nowadays, convolutional neural networks (CNNs) are probably the most efficient tools in this field[7–9], and there are efforts to use CNNs to automate colony counting.[10–12]

Our group aims to train neural networks to automate CFU estimation as well. A prerequisite for colony detection and classification with CNN is the availability of as many digital images of annotated bacterial cultures as possible. At the beginning of our work, we could not find a similar public, freely available database to use for our own CNN-based development.

The aim of creating the dataset presented here was to build a collection of digital records of bacterial cultures performed under everyday laboratory conditions on solid media. In creating such databases, the question arises as to whether the digital images should be produced under some highly controlled, standardized conditions or in a way that could presumably be produced anywhere. The former solution may obviously lead to more accurate results on a given dataset, but the latter may open up the possibility of extendibility. In creating the dataset presented here and made freely available, we chose the latter approach, using mobile phones to take 373 digital images of cultures of 24 bacterial species on solid media, annotating a total

of 57,028 bacterial colonies.

Methods

Culturing of bacterial species

Our studies have cultured 24 bacterial species of veterinary importance (Table 1). These are species whose disease processes can cause significant economic damage in farm animals, which can cause disease in companion animals, and which are important for the safety of food products. Bacterial cultures were obtained from the bacterial strain collection of the Bacteriology Laboratory, Department of Microbiology and Infectious Diseases, University of Veterinary Medicine, where the bacterial strains were stored in an ultra-low freezer at -80°C. Different media were used depending on the requirements of each bacterial species (Table 1).

Bacteria species	ID	Gram	Culturing	Agar	Required	
					NAD	CO_2
Actinobacillus equuli	sp01	-	aeorob	blood		
Actinobacillus pleuropneumoniae	sp02	-	aeorob	chocolate	+	
Aeromonas hydrophila	sp03	-	aeorob	blood		
Bacillus cereus	sp04	+	aeorob	blood		
Bibersteinia trehalosi	sp05	-	aeorob	blood		
Bordetella bronchiseptica	sp06	-	aeorob	blood		
Brucella ovis	sp07	-	aeorob	blood		+
Clostridium perfringens	sp08	+	anaerob	blood		
Corynebacterium pseudotuberculosis	sp09	+	aeorob	blood		
Erysipelothrix rhusiopathiae	sp10	+	aeorob	blood		
Escherichia coli	sp11	-	aeorob	nutrient		
Glaesserella parasuis	sp12	-	aeorob	chocolate	+	+
Klebsiella pneumoniae	sp13	-	aeorob	blood, nutrient		
Listeria monocytogenes	sp14	+	aeorob	blood		
Paenibacillus larvae	sp15	+	aeorob	blood		+
Pasteurella multocida	sp16	-	aeorob	blood		
Proteus mirabilis	sp17	-	aeorob	MacConkey		
Pseudomonas aeruginosa	sp18	-	aeorob	blood		
Rhodococcus equi	sp19	+	aeorob	blood		
Salmonella enterica	sp20	-	aeorob	nutrient		
Staphylococcus aureus	sp21	+	aeorob	blood		
Staphylococcus hyicus	sp22	+	aeorob	blood		
Streptococcus agalactiae	sp23	+	aeorob	blood		
Trueperella pyogenes	sp24	+	aeorob	blood		

Table 1. The bacterial species included in the data set. The ID column contains the unique identifier of the species, while the third column contains its Gram-staining characteristics. The culture column shows whether the bacterium requires an aerobic or anaerobic environment, and the agar column shows the medium in which it is grown. The last two columns indicate whether the species requires nicotinamide adenine dinucleotide (NAD) or CO_2 during incubation.

Several steps were necessary to obtain the bacterial cultures we later used to make digital images. On the first day, the frozen strains were inoculated onto the appropriate culture medium for the bacteria and incubated under conditions appropriate to the requirements of the bacteria. On the second day, a typical colony from the culture was inoculated onto a fresh medium and incubated. On the third day, a colony of bacteria was inoculated into tryptone soy broth (TSB) using a sterile cotton swab and incubated at 37°C for 24 hours. The cultures were then used to prepare a dilution series on a decimal basis using sterile physiological saline suspension. In the first step of the dilution (basic dilution), 0.1 ml of the initial culture was first pipetted into a test tube containing 9.9 ml sterile saline, and the suspension was thoroughly homogenized (10^{-2} dilution). Then 0.5 ml of this suspension was pipetted into a test tube containing 4.5 ml sterile physiological saline solution. This gave the 10^{-3} dilution. The latter step of the dilution was carried out up to the 10^{-6} dilution (further dilutions).

Each member of the dilution series was homogenized by vortexing for 10 seconds. Subsequently, 50 µl per dilution of the dilutions was taken from each medium and distributed over the surface of the medium using a sterile glass rod with circular movements. After a final incubation at 37°C for 24-48 hours, digital images of the Petri dishes containing the cultures were taken.

Digitalization and annotation

For the digitalization, two different mobile phones were used so that the variability of the devices could be accounted for in the data set. For the same purpose, black and white backgrounds for the dishes were used to take the photos. Care was taken to ensure that the camera on the phone was parallel to the plane of the Petri dish.

The digital images were uploaded to a server where an expert using COCO Annotator (v0.11.1)[13], drew a bounding box around each colony and labeled the identified unit with the bacterial species (Figure 1). After annotation, the COCO[14] structured JSON[15] files containing the bounding boxes and labels were downloaded and subjected to further validation steps.

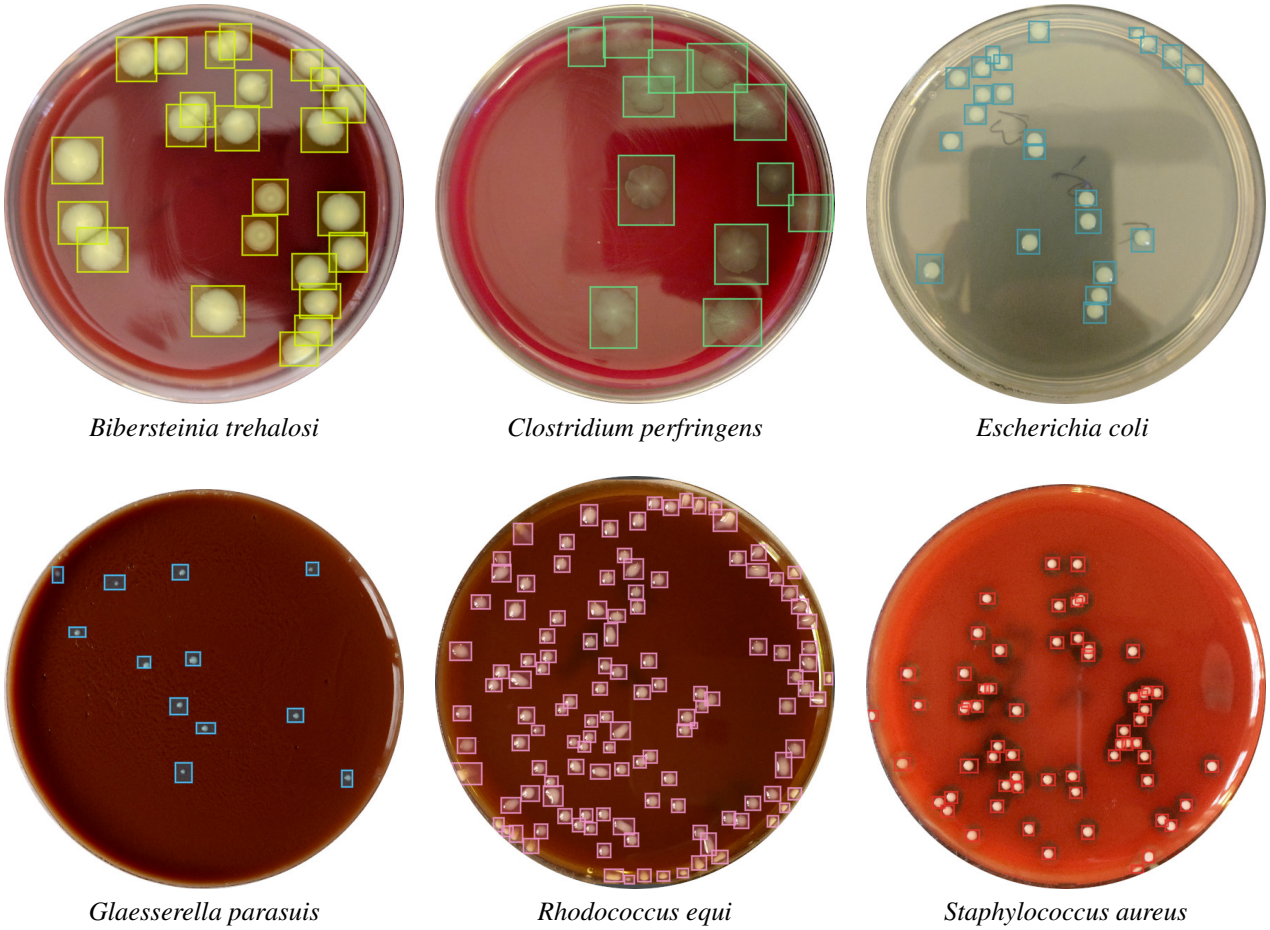

Figure 1. Six of 24 bacterial species cultures in Petri dishes with characteristic colonies annotated by bounding boxes. Each species has been cultured on the appropriate medium, e.g., *B. trehalosi*, *C. perfringens*, *R. equi* or *S. aureus* on blood agar, *G. parasuis* on chocolate agar and *E. coli* on nutrient agar.

Data Records

The number of images, annotations per bacterial species, and the distribution of the number of annotations per image are summarized in Figure 2. The curated 373 digital images of 24 bacterial species cultures are freely available in the Figshare repository (`https://doi.org/10.6084/m9.figshare.22022540`). The filenames describe their origin. The first member of the file name is the bacterial species identifier (the ID column in Table 1), and the second member is the serial number of the image associated with that species. Accordingly, the naming file `sp21_img04.jpg` is the 4th image of the *Staphylococcus aureus* cultures. In addition to the images, the repository contains one metadata and five annotation files. One line of the metadata file (`images.xls`) for a digital image contains the bacterial species ID, the file name, whether it was taken on a white or black background, and how many CFUs it contains. The `annot_COCO.json`, `annot_tab.csv`, `annot_tab.tsv`, `annot_VOC_XML.zip` and `annot_YOLO.zip` files contain the 57,028 annotation data in COCO JSON, comma-separated, tab-separated, Pascal VOC XML[16] and YOLO[17] formats respectively.

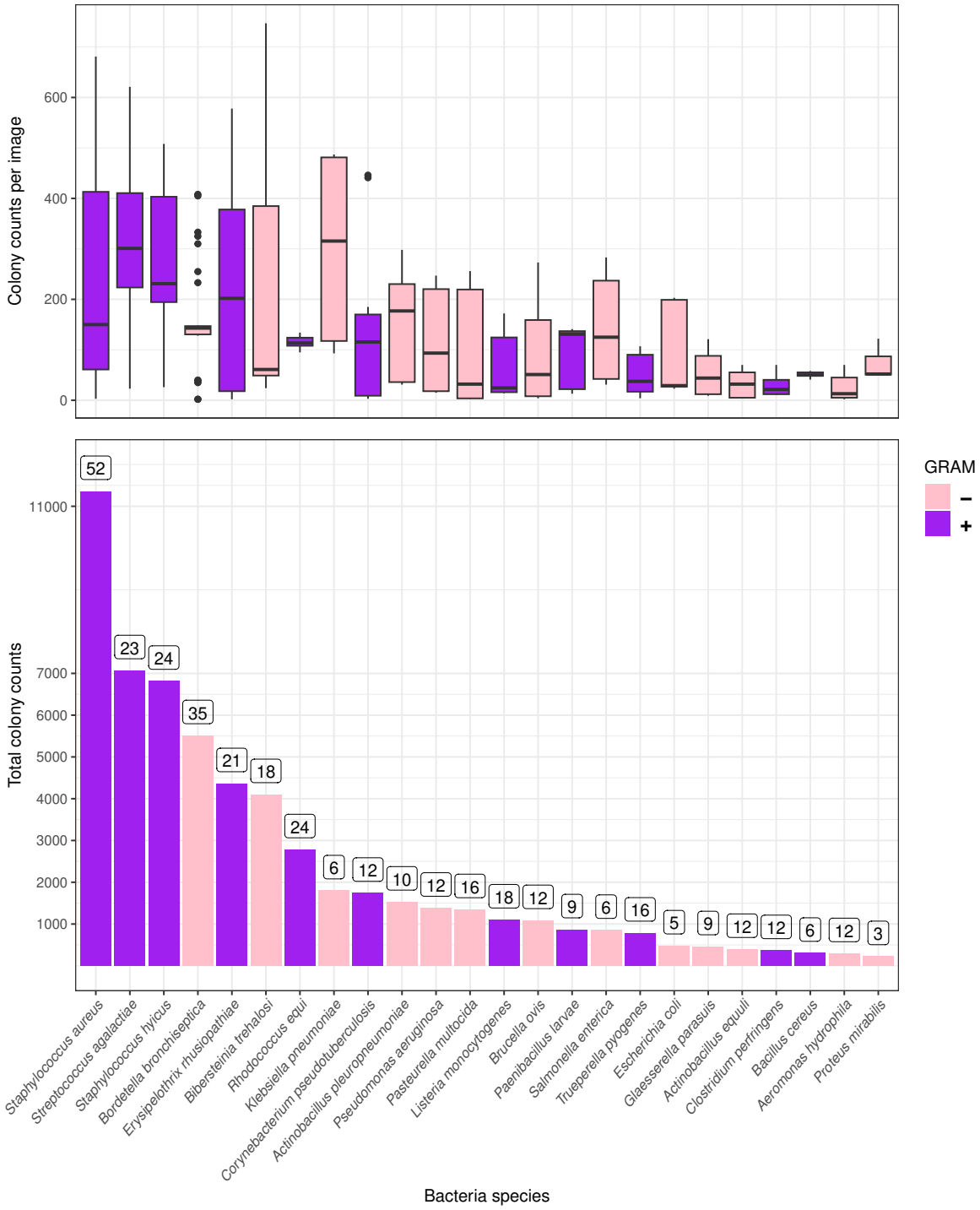

Figure 2. Distribution of colony counts by species and images. The barplots represent the total number of annotated colonies by species and the number of images belonging to the species in the dataset above the bars. The boxplots summarize the distribution of the annotations per image. The coloring of the graph shows the Gram-staining of the bacterial species.

Technical Validation

The annotated images of the bacterial cultures were curated by two experts with PhDs in bacteriology, and images they considered inappropriate were excluded from the final collection. The criteria for retaining images was whether the bacterial

colonies morphologically matched the criteria for the species completely.

The annotations exported from the COCO annotator were reviewed by another expert using the Make Sense (v1.11.0-alpha)[18] tool, and the necessary corrections were made. In some cases, two identical images of the same culture were included in the initial collection, and these redundancies were also removed in this step.

As our previous experience has shown that annotation bounding boxes exported from some annotation software can shift, especially for large numbers of annotated objects, we checked these separately. Since our CNN training designed on the dataset will be performed in the Detectron2[19] environment, we tested whether the position of the annotation bounding boxes on the images placed with Detectron2 is correct based on the COCO format JSON files generated from the CSV files exported from Make Sense. This was done using a Python script that placed the associated bounding boxes on each digital image. The resulting images were curated one by one, and in all cases, the annotation bounding box positions were found to be correct.

Usage Notes

As we have more experience in object detection and classification with Detectron2, we recommend this environment for using the data. As several other efficient solutions are available, we have placed the annotation data in the repository in various formats to facilitate wider use of the data.

We believe the dataset can be used for three types of object detection and classification tasks. The first option is to train neural networks to detect bacterial colonies separately per species. A second option is to treat colonies of 24 species with different morphologies as one class and train CNNs on the whole dataset to detect a "general colony-forming unit" type. A third option is to train the CNN on all the bacterial culture images and annotations but using the 24 classes, allowing the classification of bacterial colonies in addition to detection.

Code availability

As mentioned above, the correct position of the annotations was verified by drawing the corresponding bounding boxes on the images using Detectron2. The Python script used for this is in the file `bbox_placement_test.py`. The input annotation file for this run is a COCO JSON one. This was also generated from the tab-delimited annotation file using a Python script provided in `TSV_to_COCO.py`. Both script files are available in the Figshare repository.

References

1. McVey, D. S., Kennedy, M., Chengappa, M. M. & Wilkes, R. *Veterinary microbiology* (John Wiley & Sons, 2022), 4th edn.
2. Pau, G., Fuchs, F., Sklyar, O., Boutros, M. & Huber, W. EBImage–an R package for image processing with applications to cellular phenotypes. *Bioinformatics* **26**, 979–981 (2010).
3. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. methods* **9**, 671–675 (2012).
4. Geissmann, Q. OpenCFU, a new free and open-source software to count cell colonies and other circular objects. *PloS ONE* **8**, e54072 (2013).
5. Torelli, A., Wolf, I., Gretz, N. *et al.* AutoCellSeg: robust automatic colony forming unit (CFU)/cell analysis using adaptive image segmentation and easy-to-use post-editing techniques. *Sci. Reports* **8**, 1–10 (2018).
6. Zhang, L. Machine learning for enumeration of cell colony forming units. *Vis. Comput. for Ind. Biomed. Art* **5**, 1–8 (2022).
7. Ribli, D., Horváth, A., Unger, Z., Pollner, P. & Csabai, I. Detecting and classifying lesions in mammograms with deep learning. *Sci. Reports* **8**, 4165 (2018).
8. Kilim, O. *et al.* Physical imaging parameter variation drives domain shift. *Sci. Reports* **12**, 21302 (2022).
9. Nagy, S. Á., Kilim, O., Csabai, I., Gábor, G. & Solymosi, N. Impact evaluation of score classes and annotation regions in deep learning-based dairy cow body condition prediction. *Animals* **13**, 194 (2023).
10. Ferrari, A., Lombardi, S. & Signoroni, A. Bacterial colony counting with convolutional neural networks in digital microbiology imaging. *Pattern Recognit.* **61**, 629–640 (2017).
11. Huang, L. & Wu, T. Novel neural network application for bacterial colony classification. *Theor. Biol. Med. Model.* **15**, 1–16 (2018).
12. Beznik, T., Smyth, P., de Lannoy, G. & Lee, J. A. Deep learning to detect bacterial colonies for the production of vaccines. *Neurocomputing* **470**, 427–431 (2022).
13. Brooks, J. COCO Annotator. https://github.com/jsbroks/coco-annotator/ (2019).

14. Lin, T.-Y. *et al.* Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755 (Springer, 2014).

15. Pezoa, F., Reutter, J. L., Suarez, F., Ugarte, M. & Vrgoč, D. Foundations of JSON schema. In *Proceedings of the 25th international conference on World Wide Web*, 263–273 (2016).

16. Everingham, M. *et al.* The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vis.* **111**, 98–136 (2015).

17. Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779–788 (2016).

18. Skalski, P. Make Sense. https://github.com/SkalskiP/make-sense/ (2019).

19. Yuxin Wu and Alexander Kirillov and Francisco Massa and Wan-Yen Lo and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2 (2019).

Acknowledgements

This work has been supported by the European Union project RRF-2.3.1-21-2022-00004 within the MILAB Artificial Intelligence National Laboratory framework.

Author contributions statement

N.S. takes responsibility for the data's integrity. N.S., L.M., G.S., and I.C. conceived the concept of the work. L.M., B.F., and P.C. performed the bacteria culturing. B.F. and P.C. made the digital images. B.F. and S.Á.N. annotated the images. L.M. and G.S. curated the digital images. N.S. curated and edited the annotations. N.S., S.Á.N., B.F., and L.M. participated in the drafting of the manuscript. N.S., S.Á.N., L.M., G.S., and I.C. carried out the manuscript's critical revision for important intellectual content. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests or personal relationships that could have influenced the work reported in this paper.