scientific data

OPEN

DATA DESCRIPTOR

An annotated dataset of Gram stains from positive blood cultures

Qiaolian Yi ✉, Xiaoyan Gou, Renyuan Zhu, Xiuli Xie, Mengting Hu, Xing Wang, Tai'e Wang, Kaiwen Xu & Ying-Chun Xu ✉

Bloodstream infections (BSIs) of high morbidity and mortality are across all age groups, and urgent for accurate intervention. Gram stain interpretation of positive blood cultures (PBCs) is crucial for early diagnosing BSIs, yet this manual process is labor-intensive, time-consuming, and highly operator-dependent. Artificial intelligence (AI)-assisted microscopic interpretation of stained smears presents beneficial to microbiology diagnostics. Addressing the auto-identification of blood-culture Gram stains, this study introduces a dataset of Gram-stain smears collected in clinical practice. The dataset includes 505 microscopic images, covering up to 57 species associated with BSIs, with a total of 7528 annotations. These annotations categorized by staining characteristics and morphological features into cocci, bacilli, and fungi. We trained and validated an object detection model based on the YOLOv10 architecture on this dataset to automatically localize and classify these morphological categories in microscopic images. The publicly released dataset will help developments that utilize artificial intelligence to auto-interpretate the Gram stains from PBCs for routine clinical application.

Background & Summary

Bacterial and fungal invasions into the bloodstream, leading to bloodstream infections (BSIs), are common and critical conditions in clinical practice, with multiple studies indicating a mortality rate of over 10%[1–4]. Even in areas with abundant medical resources, sepsis and septic shock caused by BSIs are significant causes of patient mortality and contribute substantially to the economic burden of healthcare[5–7]. The risk of death significantly increased as the delay of receiving appropriate drug treatment[3,8]. Timely availability of microbiological results from positive blood cultures (PBCs) is essential to enable early pathogen-directed therapy[5]. Although molecular techniques develop rapidly, blood culture remains the reference standard and first line method in BSIs diagnosing. The classical analytical process of microbiological BSIs diagnostics follows a three-tiered reporting system, where the laboratory reports the results of the PBCs smear to the clinician immediately (Tier 1 report) and preliminary identification and susceptibility results (Tier 2 report) before the final identification and susceptibility results are reported (Tier 3 report)[9]. Once the results of the blood culture smear are obtained, clinicians are able to adjust therapy based on the staining and morphological characteristics of the pathogen[10].

The Tier 1 report is based on manual microscopic examination of smears from PBCs. Microscopic interpretation of stained smears remains labour-intensive, time-consuming, and operator-dependent. Such a subjective method can be prone to poor standardization, potentially leading to incorrect interpretations or misdiagnoses[11]. Thus, auto image analysis to identify Gram stain characteristics has great potential[12]. There are several studies on automated interpretation of blood culture Gram stains by using artificial intelligence (AI)[13,14], yet they have all been conducted with non-public data.

Pathogens causing BSIs are varied. To address the lack of publicly available, high-quality datasets and to support ongoing research in automated microbial identification, we curated a dataset comprising 505 high-resolution microscopic images from real clinical PBC smears. The dataset contains a total of 7528 annotated microbial cells, covering 57 clinically relevant BSI pathogens, encompassing Gram-positive cocci (in clusters or chains/pairs), Gram-negative rods, and fungi. Critically, all specimens originated from real patient samples processed during routine diagnostic workflows, no artificially inoculated or spiked samples were used. The images capture natural variations in smear thickness, staining intensity, and background artifacts inherent to manual slide preparation, thereby reflecting the complexity encountered in actual clinical microscopy. This makes the dataset a realistic and valuable resource for both clinical reference and algorithm development.

Department of Laboratory Medicine, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, 100730, China. ✉e-mail: yiqiaolian@pumch.cn; xycpumch@139.com

Methods

Ethics statement. The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of Peking Union Medical College Hospital (I-23PJ1906, November 2023). The committee granted a waiver of informed consent based on the following grounds: (i) the study is non-interventional and utilizes only residual clinical specimens that had already fulfilled their diagnostic purpose and were scheduled for routine disposal; (ii) no additional samples were collected for research purposes, and there was no direct contact with patients; (iii) the research poses no more than minimal risk to participants, as it involves no intervention, no disclosure of results to patients, and no financial or procedural burden; and (iv) the waiver of consent does not exempt the study from rigorous ethical review, which has been completed. In addition, all the patients were informed and signed a consent form stating that their remaining samples (blood, urine, feces, tissues, etc.) might be used for research prior to hospital admission.

To ensure confidentiality, all specimens and data were fully de-identified prior to analysis. No personally identifiable information (including names, medical record numbers, or biometric data) was collected or retained. Data are stored on secure, institution-controlled servers accessible only to authorized research personnel. Any publication or data sharing will exclude any information that could potentially identify individual participants.

Data acquisition. A total of 57 identified Gram-stained slides were collected from the clinical microbiology laboratory at Peking Union Medical College Hospital between January and May 2024. The slides of blood culture smears were generated during the course of routine clinical workup, prepared by the staff on duty. Although the origin of the data was generated from patients suspected with BSIs via their routine medical diagnosis, but no medical information was obtained from this study. In this study, only data from blood culture instruments have been collected for use, including the timing of positive blood culture reporting and the types of blood culture bottles, which are utilized solely for reference purposes. The data analyzed in this study are limited to images of positive blood culture smears.

Gram-stain smears collection. As previously mentioned, PBCs smears were obtained during routine clinical workup (Fig. 1). Once positive blood cultures were detected by the blood culture system, BACTEC™ FX system (BD Diagnostics) or BACT/ALERT®VIRTUO® (bioMérieux), positive blood culture media was aspirated with a syringe and dropped onto a glass slide to make the smear. The Gram-staining procedure was performed manually or by an automated system (PREVI Color Gram, bioMérieux). At the same time, drop of positive blood culture media was subcultured onto solid growth media. Following incubation, isolated colonies were spotted onto a target plate and subjected to matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS). Species of microorganism were confirmed by the MALDI-TOF MS analysis. Slides for microscopy image digitalization and annotation were randomly collected without pre-selection of staining quality but the species of microorganism.

Microscopy imaging and annotation. Microscopy image digitalization was carried out during the microscopic examination after Gram staining. It was obtained by a Nikon Eclipse 80i microscope and a mounted high-resolution color camera TUCSEN FL 20 with a frame rate of 5.0 fps. All slides without coverslips were imaged under a magnification of ×100 with an oil immersion objective lens (Nikon Plan Fluor 100× /1.30 oil OFN25 DIC H/N2). Fields of view with typical cell morphology under the microscope were selected and digitalized. The images were stored in jpg format with an 8-bit color depth and a resolution of 5472 × 3648 pixels, corresponding to a field of view of approximately 250 μm in diameter (based on the optical field number of 25 mm and 100× objective magnification). This yields an approximate spatial resolution of 0.046 μm per pixel.

The annotation was performed using COCO Annotator v0.11.1 (https://github.com/jsbroks/coco-annotator/), described by Makrai *et al.*[15]. Two experienced clinical microbiology technologists (≥5 years of Gram stain interpretation experience) independently annotated microbial units by drawing bounding boxes around cells or clusters based on Gram staining and morphology (Fig. 2). To ensure annotation reliability, a double-blind labeling protocol was employed, followed by an automated consistency check and expert adjudication. A customizable Image Annotation Tools was used to compare the two annotation sets for each image based on bounding box overlap; the tool is publicly available (see 'Code Availability'). Full details of the annotation validation workflow, including matching criteria, discrepancy resolution, and final consensus, are provided in the 'Technical Validation' section. The released data are in standard COCO JSON format and can be evaluated using any COCO-compatible tool.

Data Records

The dataset, comprising 505 original microscopy images of 57 clinically relevant microorganisms and their corresponding annotation files, is publicly available at the Figshare repository[16].

The image files are archived in "PBCs_microorganism_image.zip", which contains all 505 images in a flat structure. Each filename follows the format "species_abbreviation + image_number" (e.g., aba_01.jpg), where the species abbreviation (ID) links directly to the metadata in the accompanying Excel file. The technically validated annotation file is provided as "PBC_microorganism_annotation_DoubleCheck.json" in standard COCO format, containing bounding boxes that localize individual microbial units (e.g., single cells or morphologically coherent clusters) across all images in a single consolidated file. For users who prefer per-image annotation files, we also provide a complementary archive "split_annotation_DoubleCheck.zip", which contains one COCO-format JSON file for each image, named consistently with the image filename. For transparency, we also include the original annotations from the two independent annotators:

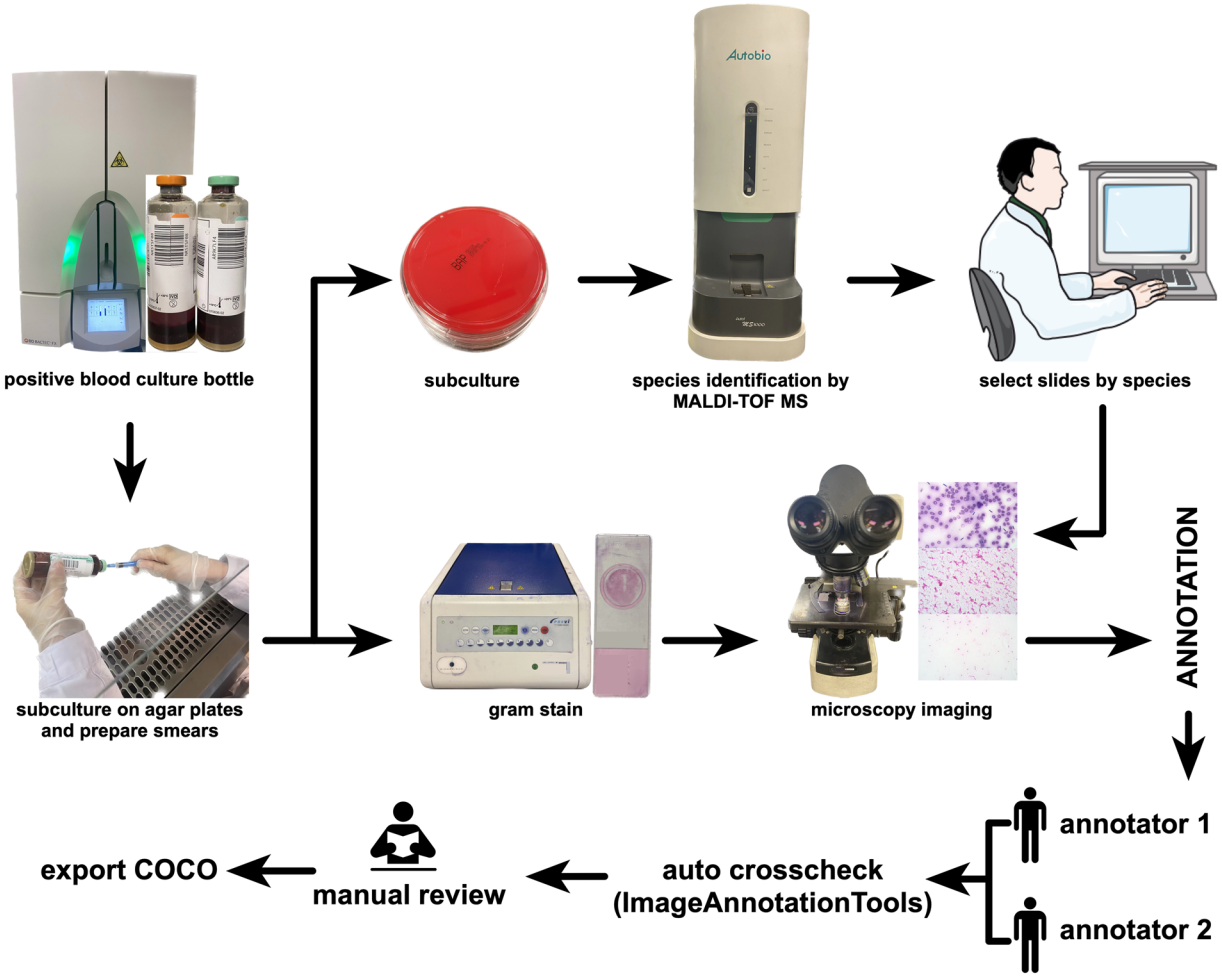

Fig. 1 Workflow of data collection and creation. Positive blood culture smears were obtained during routine clinical workup. Collected images are annotated by two independent annotators and then subjected to an automated cross-checking process followed by a manual review.

"PBC_microorganism_annotation_annotator1.json" and
"PBC_microorganism_annotation_annotator2.json".

Critically, Gram staining result (positive/negative) and morphological category (bacilli, cocci, or fungus) are not stored per bounding box in the JSON files. Instead, these microbiological attributes are provided at the species level in the accompanying file "PBCs_microorganism_information.xlsx". It contains two worksheets, species information and annotation count, that provide essential metadata for interpreting the dataset. The first worksheet (species information) lists, for each of the 57 microorganisms, its full species name, abbreviation (ID), Gram stain status, morphology type, culture vial type, time-to-positive (in hours), and the number of annotated images. The second worksheet (annotation count) offers a per-image summary, pairing each image filename with the total number of bounding box annotations it contains.

For convenience, we have also included a minimal working example script "COCO_Annotation_Visualizer.py" that reads the JSON file and displays an image with its bounding boxes, requiring only standard scientific Python libraries (Pillow, Matplotlib, pycocotools).

Technical Validation

Microorganism identification. Species of microorganism have been clinically confirmed and accurately identified. To be detailed, samples extracted from blood culture vials were then isolated with plate culture. The freshly cultured microorganism isolate was identified by MALDI-TOF MS (Autof MS 1000, Autobio Diagnostics). Only results met the criterion of score $\geq$9.0 species-level reliable identification were brought into selection of PBCs smears. The microorganism information was curated by a post-doctoral researcher in clinical microbiology, and inappropriate images were excluded in the dataset.

Data annotation validation. To minimize human error and enhance annotation consistency, we implemented a double-blind annotation and validation pipeline. Two annotators independently labelled all images without knowledge of each other's results. The customizable Image Annotation Tools used to compare the two

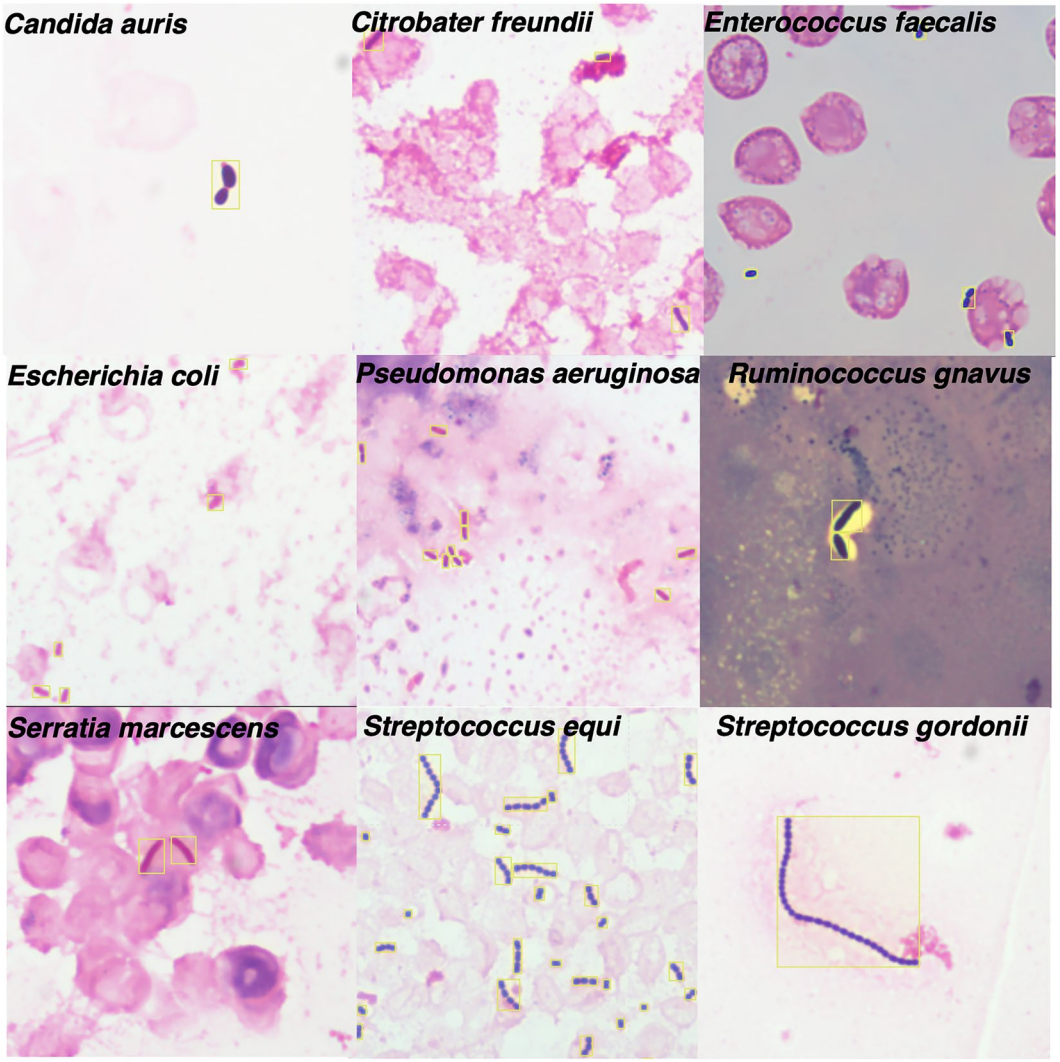

Fig. 2 Nine of 505 images with specific microbial cells annotated by bounding boxes. Each species was isolated from positive blood culture.

annotation sets for each image follow the following rule: bounding boxes from different annotators were considered a match if their Intersection over Union (IoU) exceeded 0.1, a threshold chosen to accommodate natural variability in delineating clustered or chain-forming microbes while preserving biological relevance. For matched pairs, the bounding box with the smaller area was retained to prioritize spatial precision. All unmatched annotations (IoU $\leq$ 0.1) or those with conflicting morphological labels underwent manual review by a senior microbiological morphology expert with over 20 years of frontline clinical experience, who rendered the final consensus decision. This multi-stage quality control process ensures high fidelity of the ground-truth labels used for model training and evaluation.

YOLOv10 for microorganism detection. In addition, this dataset is designed to support the development of AI algorithms for the preliminary identification of microorganisms in positive blood cultures, which is a significant step in the diagnostic workflow. We trained and validated this dataset using YOLOv10 algorithm. The training benchmark was run locally, using an NVIDIA GeForce RTX 4090, with the image resized to 1280 while maintaining the original aspect ratio. We trained the model for a total of 500 epochs, splitting the labeled data into an 8:2 training to validation ratio. During this process, we monitored several key loss indicators: 'train_box_loss' and 'val_box_loss' for bounding box accuracy, 'train_cls_loss' and 'val_cls_loss' for classifying cell types, and 'train_dfl_loss' and 'val_dfl_loss' for the distribution focal loss, which focuses on balancing the detection of cells of varying sizes and distributions. The model's performance was evaluated using metrics such as 'precision(B)', 'mAP50(B)', and 'recall(B)', where 'mAP50' denotes the mean average precision at an IoU threshold of 0.50, and 'mAP(50–95)' represents the mean average precision across a range of IoU thresholds from 0.50 to 0.95. The training and validation results are presented in Fig. 3, achieving an mAP50 of 84.6%, which demonstrated the dataset's practical utility for AI applications.

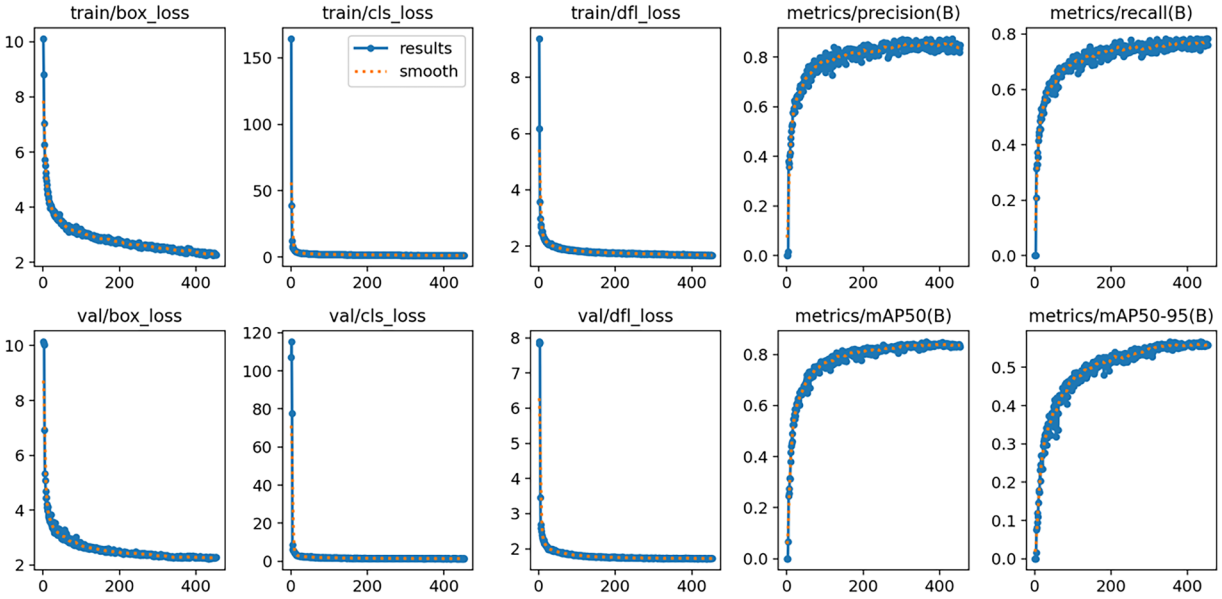

Fig. 3 Training and validation results of YOLOv10 models using the dataset of Gram stains from positive blood cultures.

Usage Notes

Dataset utilization. To utilize this dataset effectively, users can follow these steps:

Load Images: Extract the "PBCs_microorganism_image.zip" archive to access the 505 microscopy images. Each image filename (e.g., "aba_01.jpg") includes a species-specific abbreviation that links to the metadata in "PBCs_microorganism_information.xlsx".

Parse Annotations: Use any COCO-compatible tool (e.g., pycocotools in Python) to load the validated annotations using either: (i) the consolidated file "PBC_microorganism_annotation_DoubleCheck.json", or (ii) the per-image annotation files provided in "split_annotation_DoubleCheck.zip", where each JSON corresponds to a single image. Each entry provides bounding box coordinates ('bbox') and an 'image_id' corresponding to the image filename.

Retrieve Biological Attributes: Consult the first worksheet of "PBCs_microorganism_information.xlsx" to map the species abbreviation (from the image filename) to its Gram stain status, morphology type (bacilli/cocci/fungus), and other culture metadata. Since each image contains only one species, all bounding boxes within it share the same biological attributes.

Validate Annotation Counts: The second worksheet lists each image filename alongside its total number of labelled cells/clusters, which can be used to cross-check against the JSON file during data loading or debugging.

Visualize Annotations: To facilitate immediate validation and visualization, we provide a lightweight utility script, "COCO_Annotation_Visualizer.py", in the repository. This script loads a specified image and overlays its corresponding bounding box annotations using only widely available Python libraries (Pillow, Matplotlib, pycocotools).

This structure enables straightforward integration into object detection pipelines while preserving clinically relevant microbiological context.

Dataset application. We believe that the dataset can be utilized for three types of target detection and classification tasks. The first option is to treat all species as one category, training to detect and differentiate pathogens from other substances in the field of view, such as culture impurities, cellular debris, and stain residues, to identify true positive cultures and distinguish false positives. The second option is to differentiate various microbial morphologies into distinct categories for morphology classification, in order to achieve the initial goal of a Tier 1 report. The third option is to treat each species of microorganism as a separate category, mining the potential characteristic of different species of microorganisms to achieve classification at the species level. It should be noted that in the clinical practice of positive blood cultures, the primary report does not require identification of individual cells/ single cell. This dataset is primary used to identify the presence of microorganisms in the image, distinguishing them from other cells. Cell counting is not a meaningful metric for the analysis of blood culture smears.

Limitations. Limitations of the dataset or the potential constraints that may be encountered when utilizing this dataset in further research or applications are also presented as follow:

1) During the annotation process, although we have endeavored to annotate according to individual cells, there is still the possibility of a group of cells being present. At the same time, cells with incomplete edges at the edge of the image have been explicitly excluded from the annotation range. Each field of view may

have unannotated cells, but the annotated cells have all been technically checked and manually reviewed to confirm that they are the target objects.
2) In this dataset, images of each type of microorganism all come from the same slide, which means that the generalization of the same type of microorganism may be insufficient. To minimize this as much as possible, we also provide bacterial information for each slide, including its culture conditions and culture cycle.
3) Since the images are manually photographed and collected, they are subject to the limitations of optical microscope imaging. Microorganisms in the field of view may be located on different focal planes, resulting in some cells being in focus and others being out of focus. During the annotation process, both clearly outlined and blurred microorganisms have been annotated without specific distinction.

Data availability

The dataset is available at the Figshare repository[16].

Code availability

The annotation tool used for the dataset labelling is publicly available in GitHub, https://github.com/jsbroks/coco-annotator/. The customizable Image Annotation Tools used for the dataset labelling technical check is available from https://github.com/KeyOfSpectator/ImageAnnotationTools, including Double Check IoU Annotation Tool and COCO Json Merge/Split Tool.

Received: 5 May 2025; Accepted: 19 January 2026;
Published online: 23 January 2026

References

1. Jin, L. *et al*. Clinical Profile, Prognostic Factors, and Outcome Prediction in Hospitalized Patients With Bloodstream Infection: Results From a 10-Year Prospective Multicenter Study. *Front Med (Lausanne)* **8** (2021).
2. Dubourg, G., Raoult, D. & Fenollar, F. Emerging methodologies for pathogen identification in bloodstream infections: an update. *Expert Rev Mol Diagn* **19**, 161–173 (2019).
3. Adrie, C. *et al*. Attributable mortality of ICU-acquired bloodstream infections: Impact of the source, causative micro-organism, resistance profile and antimicrobial therapy. *J Infect* **74**, 131–141 (2017).
4. Ikuta, K. S. *et al*. Global mortality associated with 33 bacterial pathogens in 2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet* **400**, 2221–2248 (2022).
5. Timsit, J. F., Ruppé, E., Barbier, F., Tabah, A. & Bassetti, M. Bloodstream infections in critically ill patients: an expert statement. *Intensive Care Med* **46**, 266–284 (2020).
6. Cecconi, M., Evans, L., Levy, M. & Rhodes, A. Sepsis and septic shock. *The Lancet* **392**, 75–87 (2018).
7. Kern, W. V. & Rieg, S. Burden of bacterial bloodstream infection—a brief update on epidemiology and significance of multidrug-resistant pathogens. *Clinical Microbiology and Infection* **26**, 151–157 (2020).
8. Pien, B. C. *et al*. The clinical and prognostic importance of positive blood cultures in adults. *American Journal of Medicine* **123**, 819–828 (2010).
9. Lamy, B., Sundqvist, M. & Idelevich, E. A. Bloodstream infections – Standard and progress in pathogen diagnostics. *Clinical Microbiology and Infection* **26**, 142–150 (2020).
10. Ito, H. *et al*. The role of Gram stain in reducing broad-spectrum antibiotic use: A systematic literature review and meta-analysis. *Infect Dis Now* **53**, 104764 (2023).
11. Thomson, R. B. One small step for the Gram stain, one giant leap for clinical microbiology. *J Clin Microbiol* **54**, 1416–1417 (2016).
12. Smith, K. P. & Kirby, J. E. Image analysis and artificial intelligence in infectious disease diagnostics. *Clinical Microbiology and Infection* **26**, 1318–1323 (2020).
13. Smith, K. P., Kang, A. D. & Kirby, J. E. Automated interpretation of blood culture Gram stains by use of a deep convolutional neural network. *J Clin Microbiol* **56** (2018).
14. Walter, C. *et al*. Performance evaluation of machine-assisted interpretation of Gram stains from positive blood cultures. *J Clin Microbiol* **62** (2024).
15. Makrai, L. *et al*. Annotated dataset for deep-learning-based bacterial colony detection. *Sci Data* **10**, 497 (2023).
16. Yi, Q. *et al*. An annotated dataset of Gram stains from positive blood cultures. *Figshare*. https://doi.org/10.6084/m9.figshare.26004610

Acknowledgements

We appreciate the help from Mr. Shichun Feng for technical validation. This work has been supported by the Noncommunicable Chronic Diseases-National Science and Technology Major Project [No.2024ZD0532800], Peking Union Medical College Hospital Talent Cultivation Program-Category D [No. UHB12289], and Young Elite Scientists Sponsorship Program of the Beijing High Innovation Plan.

Author contributions

Q.Y. and Y.X. conceived the concept of the work. X.G., M.H. X.W. and T.W. performed the microorganism identification. Q.Y., X.G., R.Z. K.X. and X.X. made the digital images and annotated the images. R.Z. K.X. and X.X. curated the digital images. Q.Y. drafted the manuscript and performed the technical validation. Q.Y. and Y.X. had overarching administrative responsibility for the project. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Q.Y. or Y.-C.X.