Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Quality-Guided Subset Selection for Bacterial Microscopy Classification: A Compact Multi-Seed Benchmark on DIBaS

Krishna Sai Pokala

ok50250@umbc.edu

University of Maryland, Baltimore County https://orcid.org/0009-0001-3012-4018

Research Article

Keywords: bacterial microscopy, deep learning, image classification, computer vision, biomedical imaging, ResNet18, DIBaS, microscopy classification, image quality assessment

Posted Date: May 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9793981/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Quality-Guided Subset Selection for Bacterial Microscopy Classification: A Compact Multi-Seed Benchmark on DIBaS

Krishna Sai Pokala
pokalakrishnasai0712@gmail.com

***Abstract*—Microscopy-based bacterial classification is a compact but challenging computer vision problem because inter-class differences are visually subtle and available datasets are typically small. This paper presents a reproducible bacterial microscopy benchmark built on the Digital Images of Bacterial Species dataset using a ResNet18 backbone, explicit image-quality review, and multi-seed evaluation. The prepared dataset used in the completed benchmark contains 629 images across 30 classes. We compare a baseline training pipeline against a quality-guided subset selection arm derived from blur, contrast, and brightness cues. Across three seeds, the baseline achieved 0.9719±0.0122 accuracy and 0.9701±0.0127 macro-F1, while the quality-guided arm achieved 0.9298±0.0161 accuracy and 0.9275±0.0147 macro-F1. The strongest conclusion is methodological rather than purely performance-driven. Strict threshold-based quality control proved too restrictive in all quality-guided runs, so the final subset-selection mechanism relied on ranked class-preserving rescue. The resulting benchmark is suitable both as a transparent research artifact and as a compact case study of the trade-off between quality filtering and class coverage in small microscopy datasets.**

***Index Terms*—bacterial microscopy, image quality assessment, microscopy classification, ResNet18, subset selection**

I. Introduction

Automated bacterial image classification remains a practically relevant problem in biomedical image analysis because visual inspection is time-consuming and depends strongly on operator expertise. Recent deep learning methods have shown that convolutional networks can extract useful morphology cues from microscopy images, but dataset scale and quality variation continue to constrain straightforward generalization [1]–[3].

This project is intentionally framed as a compact benchmark rather than a state-of-the-art challenge entry. That framing is useful for two reasons. First, small microscopy datasets remain common in academic and pilot settings. Second, compact benchmarks are ideal for studying how dataset curation decisions influence performance, stability, and interpretability.

The central question of this work is whether a compact ResNet18 benchmark can classify bacterial microscopy images reliably, and whether quality-guided subset selection improves or clarifies downstream model behavior. The empirical answer is nuanced. The baseline arm outperformed the quality-guided arm in aggregate metrics, but the quality-guided arm revealed an important methodological fact: hard threshold-based quality control was unstable under class-coverage constraints and therefore required ranked class-preserving rescue in all completed runs.

The contributions of this paper are threefold:

1) a modular end-to-end benchmark pipeline for bacterial microscopy classification on DIBaS,
2) a controlled comparison between a baseline training regime and a quality-guided subset-selection arm under a multi-seed protocol,
3) and an analysis showing that threshold-based QC was too restrictive in this compact setting, requiring ranked rescue to preserve viable class-balanced subsets.

II. Related Work

Residual networks remain one of the most influential backbone families in computer vision because residual connections support stable optimization and strong transfer-learning behavior [4]. For compact biological image benchmarks, ResNet18 is a useful choice because it offers a credible vision baseline without introducing unnecessary architectural complexity.

Within bacterial imaging specifically, Zieliński et al. established DIBaS as a benchmark setting and showed that deep learning descriptors can support bacterial colony classification on a compact public dataset [1]. Later work broadened the imaging modalities and task formulations under consideration. Hallström et al. demonstrated label-free bacterial species classification from phase-contrast time-lapse microscopy, showing that richer optical and temporal signals can support strong species discrimination in controlled conditions [2].

At a broader level, Wu and Gadsden reviewed machine learning approaches to microbial classification and emphasized the importance of dataset design, feature quality, and model selection in determining reported performance [3]. Kim et al. studied rapid CNNs for Gram-stained image classification with a focus on practical inference efficiency, illustrating that compact and deployment-oriented image classifiers remain highly relevant in microbiological workflows [5]. These studies motivate the present work: rather than proposing a new architecture, we analyze the interaction between a compact residual baseline, simple QC cues, and subset-selection behavior on a small public microscopy dataset.

III. Dataset and Preprocessing

The benchmark uses the Digital Images of Bacterial Species dataset introduced by Zieliński et al. [1]. In the present

TABLE I
QC LABEL COUNTS IN THE PREPARED DATASET.

QC label	Count
blurry	498
acceptable	112
low_contrast	19

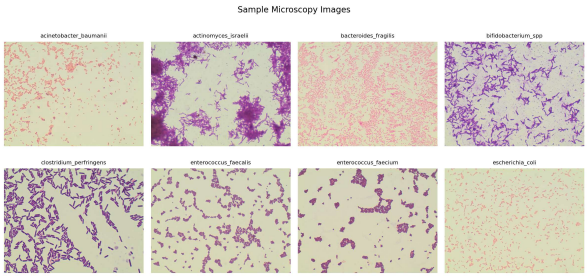


Fig. 1. Representative microscopy images from the prepared dataset.

project, *Candida albicans* is excluded so that the task remains framed as bacterial species classification. After preparation and validation, the completed dataset contains 629 images across 30 classes.

All readable images are organized into a prepared local structure before model training. A dataset manifest is then built to record simple image-quality signals for each sample. These signals include Laplacian variance as a blur proxy, grayscale standard deviation as a contrast proxy, and mean grayscale intensity as a brightness proxy. These measurements are intentionally practical and reproducible rather than domain-certified quality labels.

IV. METHOD

A. Baseline Arm

The baseline arm trains a ResNet18 classifier on the standard prepared dataset split. This arm functions as the primary benchmark reference and is intended to show how well a compact residual model can separate bacterial species in a controlled multi-seed evaluation.

B. Quality-Guided Arm

The second arm applies quality-aware subset selection before training. In code, this arm first attempts threshold-based QC on blur, contrast, and brightness signals. When direct thresholding removes too many images or disrupts class coverage, the pipeline falls back to a ranked class-preserving rescue procedure based on the composite quality ranking of available images.

This detail is central to the correct interpretation of the project. The final completed benchmark did not execute as strict threshold-only QC in any seed. Instead, all quality-guided runs used ranked rescue. The QC strategy summary was ranked_rescue: 3. Therefore, the most accurate description of the second arm is quality-guided subset selection rather than strict QC filtering.

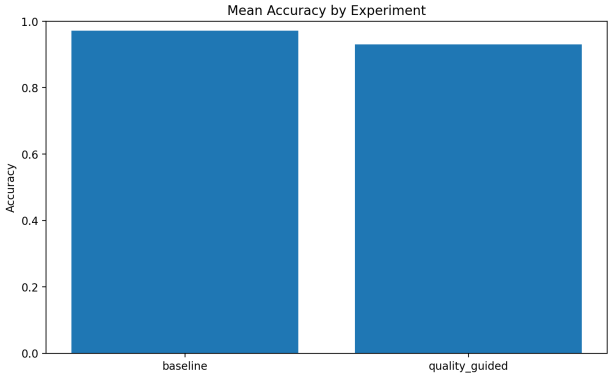


Fig. 2. Mean accuracy comparison across experiment arms.

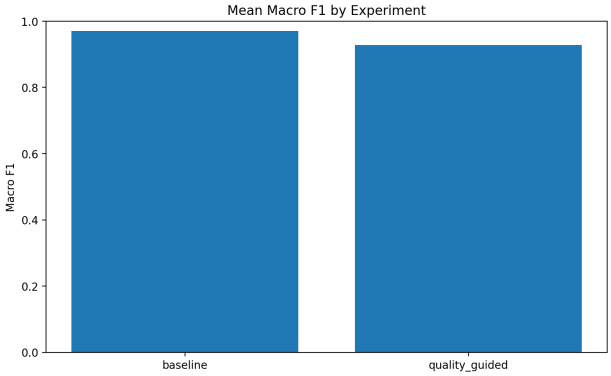


Fig. 3. Mean macro-F1 comparison across experiment arms.

C. Model and Evaluation Protocol

A ResNet18 backbone is used throughout, consistent with the benchmark's emphasis on transparency and reproducibility. Experiments are repeated for seeds 42, 123, and 999. The primary evaluation metrics are accuracy, macro-F1, and weighted-F1. Macro-F1 is especially important because it reflects class-balanced behavior more faithfully than accuracy alone in compact multi-class settings.

V. EXPERIMENTAL SETUP

The completed benchmark uses image size 224, batch size 16, 12 epochs, and a 0.70/0.15/0.15 train/validation/test split. The implementation is written in Python using PyTorch for model training and scikit-learn for evaluation and reporting [6], [7]. This configuration intentionally prioritizes clean experimental control over architectural novelty.

VI. RESULTS

The baseline arm outperformed the quality-guided arm across all aggregate metrics. When measured as quality-guided minus baseline, the deltas were -0.0421 for accuracy, -0.0426 for macro-F1, and -0.0439 for weighted-F1.

The strongest individual run recorded in the curated summaries was baseline_seed42 (seed 42, macro-F1 0.9778, accuracy 0.9789).

TABLE II
AGGREGATE MULTI-SEED BENCHMARK RESULTS.

Experiment	Acc. mean	Acc. std	Macro-F1 mean	Macro-F1 std	Weighted-F1 mean	Weighted-F1 std
Baseline	0.9719	0.0122	0.9701	0.0127	0.9717	0.0120
Quality-guided	0.9298	0.0161	0.9275	0.0147	0.9278	0.0165

TABLE III
SEED STABILITY SUMMARY.

Experiment	Acc std	Macro-F1 std	Weighted-F1 std
baseline	0.0122	0.0127	0.0120
quality_guided	0.0161	0.0147	0.0165

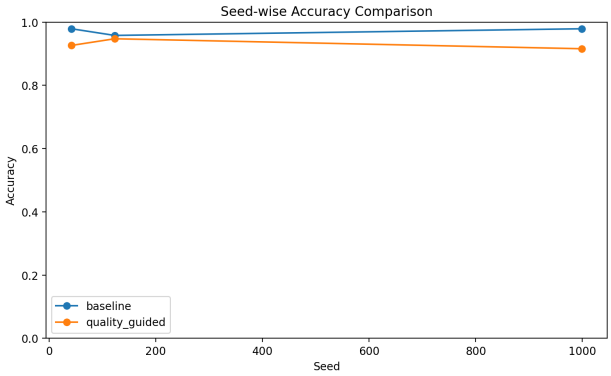


Fig. 4. Seed-wise accuracy comparison across seeds.

TABLE IV
LARGEST PREPARED CLASSES.

Class	Images
actinomyces_israelii	23
bacteroides_fragilis	23
bifidobacterium_spp	23
clostridium_perfringens	23
fusobacterium_spp	23
neisseria_gonorrhoeae	23
porphyromonas_gingivalis	23
propionibacterium_acnes	23
listeria_monocytogenes	22
veillonella_spp	22

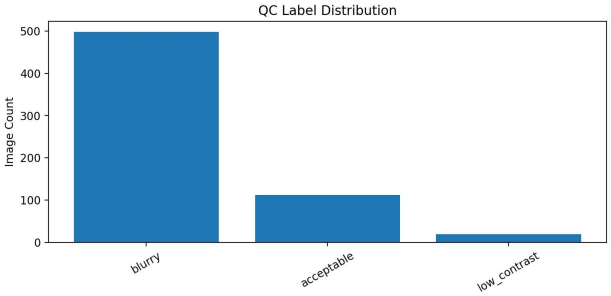


Fig. 5. QC label distribution in the prepared dataset.

VII. DISCUSSION

The most important result of this study is not that quality-guided selection improves final benchmark metrics. It does not. Instead, the most informative result is methodological. Practical QC cues were useful diagnostically, but they were too brittle as hard exclusion rules in this compact dataset. Once class coverage and subset viability were enforced, all quality-guided runs required ranked class-preserving rescue.

This finding matters because small microscopy datasets often tempt aggressive cleaning heuristics. The present benchmark suggests that such heuristics can damage representation before they improve learnability. In that sense, the quality-guided arm remains useful even though it underperforms the baseline. It makes the instability of threshold-driven filtering visible and motivates softer subset-selection logic.

VIII. LIMITATIONS

This study has several limitations. The dataset is small. The QC metrics are heuristic and not validated against expert quality labels. The benchmark uses a compact backbone rather than a broad architecture sweep. Most importantly, because all quality-guided runs used ranked rescue, this paper should not claim that threshold-only QC improved classification accuracy. The appropriate claim is narrower and more honest: quality-aware subset selection was analytically useful, but direct thresholding was unstable in this setting.

IX. CONCLUSION

This paper presented a modular bacterial microscopy benchmark with explicit image-quality review, multi-seed evaluation, and transparent reporting. The baseline outperformed the quality-guided arm in aggregate metrics, but the quality-guided analysis still yielded a meaningful methodological result. Hard QC thresholds were unstable in this compact setting, and ranked class-preserving rescue was necessary to preserve viable training subsets. This makes the project suitable both as a reproducible research portfolio artifact and as a disciplined IEEE-style and arXiv-ready case study of quality-guided subset selection on small microscopy datasets.

REFERENCES

[1] B. Zieliński, A. Plichta, K. Misztal, P. Spurek, M. Brzychczy-Włoch, and D. Ochońska, "Deep learning approach to bacterial colony classification," *PLOS ONE*, vol. 12, no. 9, p. e0184554, 2017.

[2] E. Hallström, V. Kandavalli, P. Ranefall, J. Elf, and C. Wählby, "Label-free deep learning-based species classification of bacteria imaged by phase-contrast microscopy," *PLOS Computational Biology*, vol. 19, no. 11, p. e1011181, 2023.

[3] Y. Wu and S. A. Gadsden, "Machine learning algorithms in microbial classification: A comparative analysis," *Frontiers in Artificial Intelligence*, vol. 6, p. 1200994, 2023.

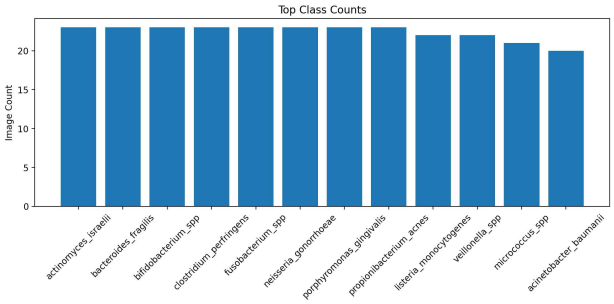


Fig. 6. Largest prepared classes by image count.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[5] H. E. Kim, M. Maros, S. R. Park *et al.*, "Rapid convolutional neural networks for gram-stained image classification at inference time on mobile devices," *Biomedicines*, vol. 10, no. 11, p. 2808, 2022.

[6] PyTorch Developers, "Pytorch," 2025, software framework.

[7] F. Pedregosa, G. Varoquaux, A. Gramfort *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.