

Classification of Gram-Stained Microscopy Images of Bacteria Cultured on Multiple Media Using Frozen and Fine-Tuned ResNet-50 Transfer Learning: A Comparative Study

Daniel Martomanggolo Wonohadidjojo^{1*}, Lidya Handayani²

¹Department of Informatics, School of Information Technology, Universitas Ciputra, Surabaya, East Java, Indonesia

²Department of Medical Microbiology, School of Medicine, Universitas Ciputra, Surabaya, East Java, Indonesia
E-mail: ^{1*}daniel.m.w@ciputra.ac.id, ²lidya.tjan@ciputra.ac.id

(Received: 1 Jun 2026, revised: 29 Jun 2026, accepted: 30 Jun 2026)

Abstract

The accurate and rapid identification of pathogenic bacteria is critical in clinical microbiology for guiding appropriate antibiotic therapy and infection control. This study proposes an automated classification system for Gram-stained bacterial microscopy images representing four clinically significant Gram-negative pathogenic species — *Enterobacter cloacae*, *Escherichia coli*, *Klebsiella pneumoniae*, and *Pseudomonas aeruginosa* — cultivated on five different culture media: Blood Agar Plate (BAP), MacConkey agar (MAC), Mueller-Hinton agar (MHA), Mueller-Hinton broth (MHB), and Nutrient agar (NA), yielding a 20-class classification problem. A primary dataset of 166 images was collected from clinical isolates using a standardized Gram staining protocol and imaged at 1000× magnification using a Leica DM500 trinocular microscope. Two ResNet-50 transfer learning strategies were compared under stratified 5-fold cross-validation: Experiment A using a frozen ResNet-50 as a fixed feature extractor, and Experiment B using a fully fine-tuned ResNet-50. Experiment B (fine-tuned) outperformed Experiment A (frozen) overall, achieving mean accuracy of 0.5832 ± 0.1109 and macro F1-score of 0.5409 ± 0.1231 compared to 0.5601 ± 0.0603 and 0.5010 ± 0.0935 respectively. The highest per-class F1-score was 0.9333 for *E. coli* on MAC and MHA in both experiments. Culture medium type is identified as a key determinant of classification difficulty, with selective and differential media yielding superior results over non-selective general-purpose media. Although fine-tuning improved performance, the relatively small dataset size and moderate overall accuracy indicate that larger-scale validation is required before clinical deployment.

Keywords: Gram Staining, Bacterial Image Classification, ResNet-50, Transfer Learning, Culture Media.

I. INTRODUCTION

Healthcare-associated infections caused by Gram-negative bacterial pathogens represent one of the most significant challenges in modern clinical medicine. Organisms such as *Enterobacter cloacae*, *Escherichia coli*, *Klebsiella pneumoniae*, and *Pseudomonas aeruginosa* are responsible for a wide spectrum of nosocomial infections including urinary tract infections, hospital-acquired pneumonia, bloodstream infections, and surgical site infections. What makes these organisms especially problematic is their capacity to acquire and disseminate resistance mechanisms notably ESBL production, AmpC beta-lactamase expression, and carbapenem-hydrolyzing enzymes that collectively narrow the spectrum of effective antimicrobial agents available for treatment [1].

In clinical microbiology laboratories, bacterial identification begins with specimen culture on appropriate growth media, followed by Gram staining and microscopic examination of the stained smear. Gram staining, first described by Hans Christian Gram in 1884 [2], remains one of the most fundamental and widely used diagnostic procedures in microbiology. By differentially staining bacterial cell walls using crystal violet, iodine mordant, alcohol decolorizer, and safranin counterstain, the technique provides rapid preliminary classification of bacteria into Gram-positive and Gram-negative categories and reveals critical information about cellular morphology, whether the organism appears as cocci (spherical), bacilli (rod-shaped), or coccobacilli, as well as arrangement patterns such as single cells, pairs, chains, or clusters.

The choice of culture medium profoundly influences the growth characteristics and visual appearance of bacterial colonies and, consequently, the morphological features observable in Gram-stained smears prepared from those cultures. Selective and differential media such as MacConkey agar (MAC) promote the growth of Gram negative while inhibiting Gram positive organisms, and produce color reactions based on metabolic activity (e.g., lactose fermentation) that provide additional discriminating features. Non-selective media such as Blood Agar Plate (BAP) and Nutrient Agar (NA) support a wide range of organisms without differential reactions, potentially resulting in less visually discriminative growth pattern across different species.

Despite the clinical importance of accurate and timely bacterial identification, the conventional workflow depends heavily on the expertise and experience of trained medical laboratory scientists. Manual examination of Gram-stained smears is inherently subjective and can be inconsistent between observers. In resource-limited settings, including many regional hospitals and clinical laboratories in developing countries, there may be insufficient numbers of trained personnel to meet the volume of diagnostic tests required. Computational tools that can reliably identify bacterial species from Gram-stained microscopy images would therefore serve as a practical solution: they could extend the diagnostic reach of understaffed laboratories, shorten the time from specimen collection to treatment decision, and reduce the inconsistency that arises from reliance on individual examiner expertise [3].

Deep learning, and convolutional neural networks (CNNs) in particular, has achieved remarkable performance in image classification tasks across numerous medical domains including histopathology, radiology, dermatology, and ophthalmology [4]. Transfer learning, the technique of repurposing a deep neural network pre-trained on a large general-purpose dataset such as ImageNet for a new domain-specific classification task, has been shown to be particularly effective when the target domain has limited labeled data, a situation that is common in specialized medical and microbiological imaging [5]. ResNet-50, a 50-layer deep residual network, is one of the most widely adopted architectures for transfer learning in medical image classification, owing to its residual connections that mitigate the vanishing gradient problem and enable effective training of very deep architectures [6].

Previous studies applying deep learning to bacterial image classification have primarily focused on either colony plate photography or Gram-stained microscopy images on a single culture medium. The systematic investigation of classification performance across multiple culture media types using the same bacterial species on each medium has received little attention. Understanding how different culture media induce subtle morphological variations in Gram-stained bacteria is important for evaluating the robustness and generalizability of automated diagnostic systems.

Recent studies have demonstrated the growing role of AI and deep learning in clinical microbiology, including

automated colony counting, antimicrobial susceptibility testing interpretation, and pathogen detection from microscopy images [7][8]. A comprehensive review by Mairi et al. [7] highlighted that deep learning and CNNs are increasingly being deployed for enhancing pathogen detection and diagnostic imaging in bacteriology. The 2024 ASM Clinical Microbiology Open proceedings further underscored the clinical community's anticipation of widespread AI implementation in microbiology laboratories, particularly for image analysis and quality monitoring [9]. These developments confirm the clinical relevance of the automated bacterial image classification approach presented in this study.

It is worth noting that the substantially higher accuracies reported in several of the studies cited above were obtained under conditions that differ markedly from those of the present study: those datasets typically contained several hundred to several thousand images per class, used a single culture medium rather than five, and in some cases targeted broader genus-level rather than fine-grained species-medium discrimination. The substantially smaller sample size per class in the present study (8–9 images), combined with the more granular 20-class species-medium formulation, represents a considerably more data-constrained and fine-grained classification problem, which is the primary factor underlying the lower absolute accuracy reported here relative to these prior studies. This study addresses this gap by presenting a comparative evaluation of two ResNet-50 transfer learning strategies, frozen feature extraction (Experiment A) and full fine-tuning (Experiment B), for the 20-class classification of Gram-stained bacterial microscopy images spanning four species and five culture media types. Results are evaluated using stratified 5-fold cross-validation on a primary dataset of 166 images collected from clinical isolates at a clinical microbiology laboratory in Surabaya, Indonesia. The use of 5-fold cross-validation rather than a single train/test split ensures statistically reliable and generalizable performance estimates on this small dataset.

The remainder of this paper is organized as follows. Section II describes the research methodology including specimen preparation, Gram staining protocol, image acquisition, and network design. Section III presents the experimental results and discussion. Section IV states the conclusions and future research directions.

The application of machine learning and deep learning to bacterial image analysis has developed along two primary visual paradigms: analysis of colony plate photographs and analysis of Gram-stained microscopy images. Early work in automated bacterial identification used traditional machine learning approaches based on hand-crafted feature descriptors including color histograms, local binary patterns (LBP), Gabor texture features, and morphological shape descriptors extracted from microscopy images. While these approaches demonstrated feasibility on small controlled datasets, they require substantial feature engineering effort and do not generalize well to variations in staining quality, imaging conditions, or colony density.

The introduction of the DIBaS (Digital Image of Bacterial Species) dataset by Zielinski et al. [10] provided the first large-scale benchmark for Gram-stained bacterial microscopy image classification, comprising images of 33 bacterial genera. Their evaluation of multiple deep CNN architectures demonstrated that learned feature representations substantially outperformed hand-crafted features across all tested genera. This work established the feasibility of deep learning for Gram-stained bacterial image classification and provided a widely used reference dataset for subsequent studies.

More recent studies from 2024 onwards have pushed the boundaries of automated bacterial classification with newer architectures and larger datasets. Visitsattapong et al. [11] proposed a Big Transfer (BiT) model combined with graph Laplacian-based data cleaning for the DIBaS dataset, achieving 99.11% accuracy, 99.31% precision, and 99.06% F1-score, demonstrating that general visual representation learning can be effectively transferred to bacterial microscopy image classification. Abougair et al. [12] employed transfer learning with GoogLeNet and AlexNet on a dataset of 2000 images spanning 12 pathogenic bacterial species, further confirming the versatility of transfer learning strategies across different architectures and species sets. Image processing techniques have also been applied to antimicrobial susceptibility testing interpretation [13].

The emergence of Vision Transformer (ViT) architectures has opened new avenues for bacterial image classification. Shwetha et al. [14] proposed a ViT-based model for Gram-stained direct smear images, evaluating it on both private and public datasets. Their ViT-B_16 model achieved 96% accuracy, outperforming pre-trained CNN baselines including VGG16 and EfficientNetB0, suggesting that self-attention mechanisms can capture the long-range morphological dependencies in bacterial smear images that convolutional filters may miss. Similarly, Lee et al. [15] applied Vision Transformers to alignment-free bacterial pathogen identification from high-resolution microscopy, achieving robust performance across multiple species. These recent results indicate that future work on bacterial image classification should explore transformer-based architectures as alternatives to ResNet-based models.

A multi-task deep learning framework for clinical bacterial microscopy was proposed by Chin et al. [16], integrating object detection models (YOLOv4, EfficientDet, SSD-MobileNetV2) with classification for simultaneous bacteria detection and species identification from microscopic images. Their study highlighted that the combination of detection and classification in a single pipeline improves both efficiency and accuracy for clinical laboratory workflows. This multi-task paradigm represents a promising direction for extending the single-task classification approach of the present study toward more complete automation of the microbiological examination process.

Transfer learning with pre-trained CNN architectures has been extensively explored for bacterial classification. Talo [17] reported classification accuracy exceeding 99% on the DIBaS dataset using ResNet-based transfer learning with fine-

tuning. Similarly, several studies have applied VGG-16 [18], AlexNet, and Inception-v3 [19] to bacterial colony plate image classification, achieving high accuracy on proprietary laboratory datasets. However, these high accuracy figures are typically reported for single-medium or pooled-medium datasets, and the impact of culture medium type on classification performance has not been systematically analyzed.

In the broader medical image classification literature, transfer learning from ImageNet has proven effective across domains that are visually distant from natural images, including histopathology [20], retinal fundus photography, and X-ray analysis. Yosinski et al. [21] demonstrated through systematic layer-by-layer analysis that ImageNet features are transferable to new domains, with early network layers learning universal low-level features (edges, textures, colors) that generalize broadly, while later layers learn progressively more domain-specific representations. This understanding motivates the comparison between frozen feature extraction (which preserves all transferred representations) and fine-tuning (which adapts the entire network to the target domain).

The question of whether to freeze or fine-tune transferred layers is an important practical decision that depends on dataset size, domain distance from ImageNet, and computational constraints. For very small datasets, fine-tuning all layers risks overfitting to the limited training examples, frozen feature extraction may be more stable. For moderately sized datasets in domains visually distinct from ImageNet, fine-tuning typically outperforms frozen extraction. The present study contributes to this literature by providing empirical evidence for both strategies specifically in the context of Gram-stained bacterial microscopy images on multiple culture media types.

To the best of the authors' knowledge, no previous study has systematically evaluated ResNet-50 transfer learning performance across five different culture media types for Gram-negative bacterial species using Gram-stained microscopy images. This study fills that gap and provides a methodologically rigorous baseline using 5-fold cross-validation, in contrast to several existing studies that report results from single train/test splits which may be subject to high variance on small datasets [22].

II. RESEARCH METHODOLOGY

A. Bacterial Specimens and Culture Preparation

All bacterial isolates used in this study were obtained from clinical specimens processed at a clinical microbiology laboratory in Surabaya, Indonesia. The four Gram-negative species selected: *E. cloacae*, *E. coli*, *K. pneumoniae*, and *P. aeruginosa* are among the most clinically significant nosocomial pathogens, each possessing well-documented morphological and growth characteristics on standard culture media [23]. Four Gram-negative pathogenic species were selected based on their clinical significance in nosocomial infection: *Enterobacter cloacae*, *Escherichia coli*, *Klebsiella pneumoniae*, and *Pseudomonas aeruginosa*. All bacterial

isolates were obtained from patient specimens and identified using the VITEK automated microbial identification system prior to use in this study.

Each bacterial species was inoculated onto five growth media using a sterile inoculation loop. For solid media Nutrient Agar (NA), Mueller-Hinton Agar (MHA), and MacConkey Agar (MAC), a standard streaking technique was employed to achieve isolated colonies. For liquid media Mueller-Hinton Broth (MHB), direct inoculation was performed. Blood Agar Plate (BAP) containing 5% sheep blood was used as the primary isolation medium. All inoculated media were incubated at 37°C for 24 hours under aerobic conditions. Growth on solid media was confirmed by the presence of bacterial colonies on the agar surface; growth in liquid media was confirmed by turbidity change from clear to cloudy.

B. Gram Staining Protocol

Following incubation, bacterial smears were prepared from each culture on clean glass slides. For bacteria grown on solid media, one drop of sterile distilled water was placed on the slide, and 1–2 colonies were transferred using a sterile loop to create a thin, uniform bacterial suspension. For bacteria grown in liquid media, one drop of the culture suspension was placed directly on the slide. All smears were spread thinly across the slide surface to ensure a monolayer of bacteria for optimal microscopic examination, air-dried at room temperature, and heat-fixed by passing the slide over a Bunsen burner flame three times.

The Gram staining procedure followed the standardized four-step protocol summarized in Table 1. All reagents were sourced from Laboratorium Bio Analitika, Surabaya, Indonesia. The procedure was applied consistently to all smears across all media types and bacterial species to minimize batch-to-batch staining variability.

Table 1. Gram Staining Protocol

Step	Procedure	Duration
1	Apply 2% Crystal Violet solution to smear	1 minute
2	Discard Crystal Violet; rinse with running water	—
3	Apply Lugol's iodine solution	1 minute
4	Discard Lugol; rinse with running water	—
5	Apply Acetone/Alcohol decolorizer	20 seconds
6	Discard decolorizer; rinse with running water	—
7	Apply 0.25% Safranin counterstain	1 minute
8	Discard Safranin; rinse with running water	—
9	Air-dry the smear at room temperature	—

C. Microscopy and Image Acquisition

Gram-stained smears were examined under a Leica DM500 trinocular microscope (Leica Microsystems GmbH, Wetzlar, Germany) at a total magnification of 1000× (100× oil immersion objective lens combined with 10× eyepiece). One drop of immersion oil was applied to each prepared smear prior to examination under the oil immersion objective to achieve the required optical coupling and resolution at this magnification. Images were captured through the trinocular port and saved as JPEG files at a resolution of 1920 × 1440 pixels under consistent illumination and focus conditions.

A total of 166 images were collected, distributed across 20 class conditions (four species × five media types). Table 2 presents the image distribution per species and medium type. The slight imbalance across some class combinations (8–9 images per class) reflects natural variation in image quality and availability during the data collection period.

Figure 1 presents representative samples from the dataset, illustrating the visual diversity across species and culture media types. The pink-to-red staining of Gram-negative bacilli is visible across all classes, while differences in colony density, cellular arrangement, and background characteristics reflect the influence of culture medium on bacterial morphology in Gram-stained preparations.

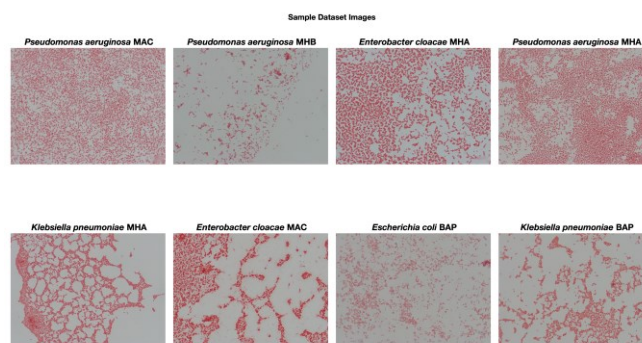


Figure 1. Representative Samples From the Dataset: Eight Gram-Stained Bacterial Microscopy Images From Two Specimen Rows, Showing Visual Variation Across Species, and Culture Media Types.

Table 2. Distribution of Gram-Stained Images per Species and Culture Medium

Species	BAP	MAC	MHA	MHB	NA	Total
<i>Enterobacter cloacae</i>	9	8	8	8	8	41
<i>Escherichia coli</i>	8	9	9	8	8	42
<i>Klebsiella pneumoniae</i>	8	9	8	9	8	42
<i>Pseudomonas aeruginosa</i>	8	8	8	9	8	41
Total	33	34	33	34	32	166

BAP: Blood Agar Plate (5% sheep blood); MAC: MacConkey Agar; MHA: Mueller-Hinton Agar; MHB: Mueller-Hinton Broth; NA: Nutrient Agar

D. Image Preprocessing

Each acquired image contained a scale bar annotation (showing "20 μm ") embedded in the lower-right corner by the Leica DM500 imaging software. This non-biological artifact, if retained during training, could serve as a spurious classification feature allowing the model to learn the presence of the scale bar rather than biologically meaningful bacterial morphology. The scale bar was therefore systematically removed from all images prior to any machine learning processing using the TELEA inpainting algorithm [24] implemented in OpenCV. This algorithm reconstructs the artifact region by propagating texture information from surrounding pixels inward along isophote directions. The removal was performed using a custom Python script that automatically detected the white pixel cluster in the lower-right region of each image and applied inpainting with a radius of 20 pixels.

Following scale bar removal, all images were resized from 1920×1440 pixels to $224 \times 224 \times 3$ pixels as required by the ResNet-50 input layer. No additional preprocessing steps such as histogram equalization or contrast normalization were applied at this stage, as the ResNet-50 architecture incorporates batch normalization layers that perform implicit input normalization during the forward pass.

E. Data Augmentation

Given the small size of the primary dataset, online augmentation was incorporated as a regularization strategy to expose the model to greater intra-class visual diversity during training. All augmentation operations were executed on-the-fly at batch generation time within each fold's training phase using MATLAB's `imageDataAugmenter`, and were strictly confined to the training subset. The validation and test subsets underwent only the mandatory resizing step to 224×224 pixels, with no augmentation applied. The following augmentation operations were applied: random horizontal reflection; random vertical reflection; random rotation uniformly sampled from $[-90^\circ, +90^\circ]$; random horizontal translation uniformly sampled from $[-20, +20]$ pixels; random vertical translation uniformly sampled from $[-20, +20]$ pixels; random horizontal scaling uniformly sampled from $[0.9, 1.1]$; and random vertical scaling uniformly sampled from $[0.9, 1.1]$. These operations were selected to simulate realistic variations in smear orientation, positioning, and magnification that may occur across different imaging sessions.

F. ResNet-50 Architecture and Modification

Deep learning architectures are built on hierarchical feature representations learned directly from raw data through multiple layers of non-linear transformations [25]. ResNet-50 is a 50-layer deep convolutional neural network introduced by He et al. [6] as part of the deep residual learning framework. The architecture is organized into one initial convolutional stem followed by four residual stages containing 3, 4, 6, and 3 bottleneck residual blocks respectively, followed by global average pooling and a fully connected classification layer. Each bottleneck block consists of three convolutional layers (1×1 , 3×3 , 1×1) with batch normalization and ReLU

activations, and a shortcut connection that adds the block input directly to its output. This residual (skip) connection allows gradients to flow directly through the network without passing through all layer transformations, addressing the vanishing gradient problem that limits the effective depth of conventional deep networks.

ResNet-50 was pre-trained on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) dataset comprising 1.2 million training images across 1000 object categories [26]. The pre-trained network learns a rich hierarchy of visual features: early layers detect low-level features (edges, color gradients, simple textures) that are broadly generalizable across visual domains, while deeper layers encode progressively more abstract and ImageNet-specific representations. For transfer learning to the bacterial microscopy image domain, the three final layers of the pre-trained network, the 1000-unit fully connected layer (fc1000), the softmax layer (fc1000_softmax), and the classification output layer (ClassificationLayer_fc1000), were replaced with a new 20-unit fully connected layer, a new softmax layer, and a new 20-class classification output layer corresponding to the 20 species-medium class combinations. The new fully connected layer was initialized with random weights and assigned weight and bias learn rate factors of 10, creating a differential learning rate that accelerates learning in the new task-specific layer while allowing more conservative, gradual updates in the transferred feature extraction layers.

The explainability of deep learning models in medical image classification has emerged as a critical requirement for clinical adoption. Chaddad et al. [27] reviewed generalizable and explainable deep learning methods for medical imaging, emphasizing that gradient-based visualization techniques such as Grad-CAM [28] are among the most widely adopted approaches for providing visual explanations of model decisions. While the present study does not implement explainability methods, their application to the ResNet-50 models trained here is identified as a priority for future work, as understanding which morphological features of Gram-stained smears drive classification decisions would be essential for building clinical trust and validating the model against microbiologist judgment. This limitation is particularly salient given the moderate accuracy achieved in the present study (approximately 58%): without Grad-CAM or a comparable visualization technique, it cannot be confirmed whether the trained models are attending to genuine bacterial cellular structures or are instead partly responding to background characteristics of the culture medium visible in the Gram-stained smear, such as residual staining artifacts or substrate texture. This concern is revisited and emphasized further as a future work priority in Section IV.

G. Experimental Configurations

Two experimental configurations were implemented to compare different transfer learning strategies under identical training conditions:

Experiment A: Frozen ResNet-50 (Feature Extraction): All pre-trained ResNet-50 layers were frozen by setting their weight and bias learn rate factors to zero. Only the new 20-unit

fully connected layer was trainable. This configuration uses ResNet-50 purely as a fixed visual feature extractor, mapping each input image to a 2048-dimensional feature vector via global average pooling, which the trainable fully connected layer then maps to class probabilities. The assumption is that ImageNet-learned features are sufficiently general to discriminate between bacterial classes without requiring domain-specific adaptation.

Experiment B: Fine-Tuned ResNet-50 (Full Transfer Learning): All layers were trainable. The pre-trained layers were assigned an initial learning rate of 1×10^{-4} , while the new fully connected layer trained at ten times this rate (1×10^{-3}). This approach allows the entire network to adapt its feature representations to the Gram-stained bacterial microscopy image domain while leveraging the general visual features already encoded in the pre-trained weights as a starting point, reducing the risk of divergence during training.

Both experiments used identical training hyperparameters: Stochastic Gradient Descent with Momentum (SGDM) optimizer with momentum 0.9, mini-batch size 16, maximum 30 training epochs, initial learning rate 1×10^{-4} , and early stopping with a validation patience of 7 (training was terminated if validation loss did not improve over 7 consecutive validation frequency checks, where validation was performed every 3 iterations). All experiments were implemented in MATLAB R2022b using the Deep Learning Toolbox.

H. Evaluation Methodology and Performance Metrics

To make full use of the limited dataset while producing statistically robust performance estimates, stratified 5-fold cross-validation was adopted as the evaluation framework. This approach partitions the data into five complementary subsets such that every image serves as a test sample exactly once, with class proportions preserved in each partition. In each fold, 80% of the data (approximately 133 images) was allocated to training and validation, and 20% (approximately 33 images) was held out as the test set. Within the training+validation portion of each fold, 85% was used for training (approximately 113 images) and 15% for validation (approximately 20 images). The cvpartition function with stratified splitting was used to ensure proportional class representation in each fold's test set. The random seed was fixed at 42 to ensure reproducibility.

Four performance metrics were computed for each fold and reported as mean $\pm$ standard deviation across the five folds. Overall accuracy measures the fraction of correctly classified test images. Macro-averaged precision, recall, and F1-score treat each class equally regardless of its sample size, making them appropriate for this dataset where class sizes are approximately equal but small. For class i , precision is defined as $TP_i / (TP_i + FP_i)$, recall as $TP_i / (TP_i + FN_i)$, and F1-score as the harmonic mean of precision and recall: $2 \times (\text{precision}_i \times \text{recall}_i) / (\text{precision}_i + \text{recall}_i)$. Macro-averaged values are computed by averaging the per-class metrics over all 20 classes. It is acknowledged that the dataset size of 166 images across 20 classes is small relative to typical deep learning benchmarks; this constraint, along with its

implications for model generalizability, is discussed further in Section III.E and revisited as a priority for future work in Section IV. To contextualize the suitability of ResNet-50 for this task, a paired-samples statistical comparison between Experiment A and Experiment B, as well as a benchmark comparison against a SqueezeNet baseline trained under identical cross-validation conditions, are reported in Section III.F.

III. RESULTS AND DISCUSSION

A. Evaluation Methodology and Performance Metrics

Table 3 presents the mean and standard deviation of all four performance metrics across the five cross-validation folds for both experimental configurations. Experiment B (fine-tuned ResNet-50) achieved higher mean values across all metrics compared to Experiment A (frozen ResNet-50). The fine-tuned model achieved a mean accuracy of 0.5832 compared to 0.5601 for the frozen model, a difference of 0.0231. The mean macro F1-score was 0.5409 for Experiment B versus 0.5010 for Experiment A, a difference of 0.0399. Mean macro precision and recall also favored Experiment B, with differences of 0.0378 and 0.0250 respectively.

Table 3. Mean $\pm$ Standard Deviation of Classification Metrics (5-Fold CV)

Metric	Exp A Mean	Exp A Std	Exp B Mean	Exp B Std
Accuracy	0.5601	0.0603	0.5832	0.1109
Macro Precision	0.5090	0.1239	0.5468	0.1323
Macro Recall	0.5600	0.0742	0.5850	0.1098
Macro F1-Score	0.5010	0.0935	0.5409	0.1231

Exp A: Frozen ResNet-50 (Feature Extraction); Exp B: Fine-tuned ResNet-50 (Transfer Learning)

Notwithstanding its lower mean performance, Experiment A exhibited markedly greater consistency across folds. Its accuracy standard deviation of 0.0603 was roughly half that of Experiment B (0.1109), indicating that freezing the pre-trained weights produces more predictable results from fold to fold when training data are scarce. Similarly, standard deviations for macro precision (0.1239 vs 0.1323), recall (0.0742 vs 0.1098), and F1-score (0.0935 vs 0.1231) were all higher for Experiment B. This indicates that while fine-tuning generally produces higher mean performance, it is more sensitive to the specific composition of each fold's training set when training samples are very limited. The training convergence behavior of both experiments is illustrated in Figures 2 and 3, which show the accuracy and loss curves for Experiment A (Frozen ResNet-50) and Experiment B (Fine-tuned ResNet-50) respectively, using Fold 5 as a representative example. Both figures confirm that training converged progressively over iterations with loss decreasing consistently, and that early stopping functioned correctly by halting training when validation performance ceased to improve.

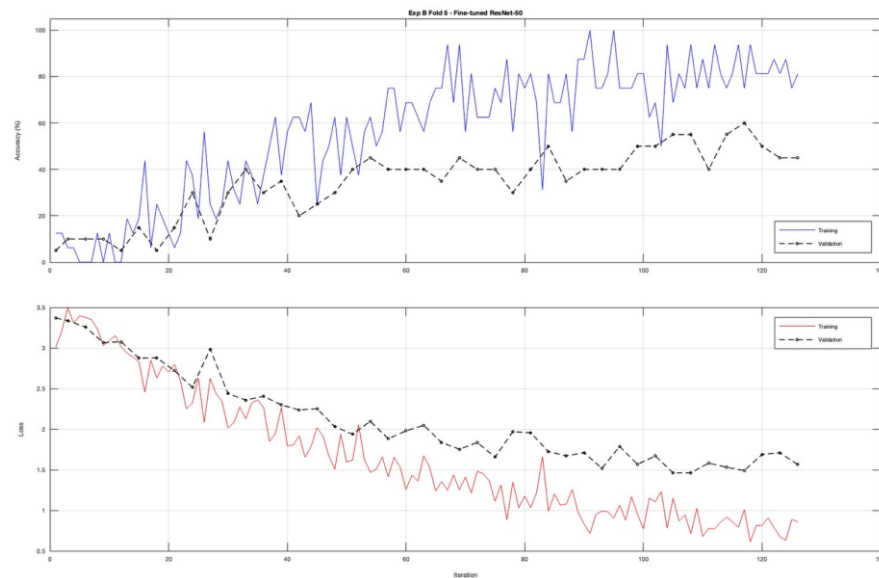


Figure 2. Training Accuracy and Loss Curves of Experiment A (Frozen ResNet-50, Representative Fold 5).

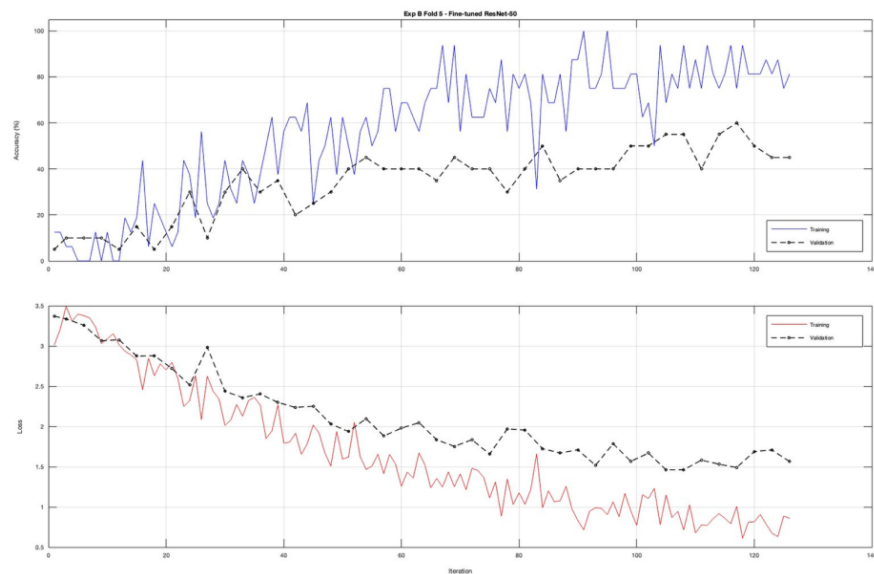


Figure 3. Training Accuracy and Loss Curves of Experiment B (Fine-tuned ResNet-50, Representative Fold 5).

Per-Fold Performance Analysis

Table 4 presents the per-fold breakdown of accuracy, F1-score, and precision for both experiments. Experiment B outperformed Experiment A in three of the five folds (Folds 1, 2, and 5), while Experiment A performed better in Folds 3 and 4. The best single-fold result across the entire study was achieved by Experiment B in Fold 2, with an accuracy of 0.7647 and macro F1-score of 0.7067. The weakest fold for both experiments was Fold 5, where Experiment A achieved an accuracy of only 0.4545 (F1-score 0.3517), indicating that the particular combination of images in that fold's test set was especially challenging for the frozen feature extractor.

Table 4. Per-Fold Performance Breakdown for Experiment A and Experiment B

Fold	Acc-A	F1-A	Prec-A	Acc-B	F1-B	Prec-B
1	0.5758	0.5610	0.5433	0.6061	0.6295	0.6467
2	0.5882	0.5402	0.5825	0.7647	0.7067	0.7250
3	0.6061	0.5821	0.6317	0.5455	0.5050	0.4917
4	0.5758	0.4700	0.4750	0.4848	0.4400	0.4500
5	0.4545	0.3517	0.3125	0.5152	0.4233	0.4208
Mean	0.5601	0.5010	0.5090	0.5832	0.5409	0.5468
Std	0.0603	0.0935	0.1239	0.1109	0.1231	0.1323

Acc: Overall Accuracy; F1: Macro F1-Score; Prec: Macro Precision

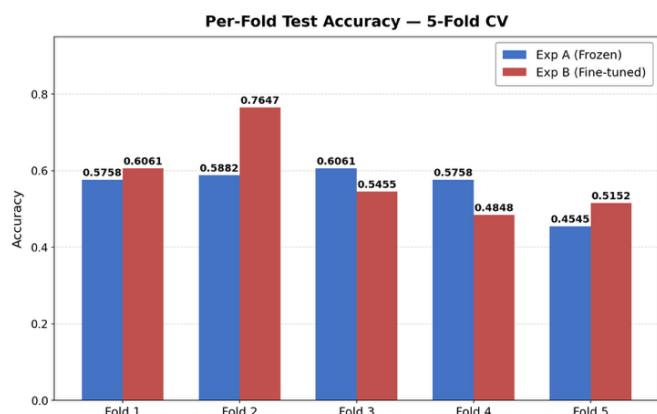


Figure 4. Per-fold Test Accuracy (%) for Experiment A (Frozen) and Experiment B (Fine-tuned) Across All 5 Folds.

The variability across folds ranging from 0.4545 to 0.6061 in accuracy for Experiment A and from 0.4848 to 0.7647 for Experiment B, highlights the statistical instability that results from having only approximately 5–6 training images per class per fold. This degree of variability underscores why single-split reporting is inadequate for small datasets: depending on which samples happen to fall in the test set, a one-off evaluation could overstate or understate the model's true average capability by a considerable margin, making the mean-and-variance summary from cross-validation indispensable. Figure 4 provides a visual comparison of per-fold test accuracy for both experiments across all five folds, clearly illustrating the higher variability of Experiment B (particularly Fold 2 reaching 76.5% versus Fold 4 dropping to 48.5%) compared to the more stable performance of Experiment A (ranging from 45.5% to 60.6%).

B. Per-Class Performance Analysis

Table 5 presents the per-class mean F1-score and precision for all 20 class combinations across both experiments. The results reveal substantial heterogeneity in classification performance across class combinations, driven primarily by the culture medium type.

Table 5. Per-Class Mean F1-Score and Precision for All 20 Classes (5-Fold CV)

Class	F1-A	F1-B	Prec-A	Prec-B
<i>E. coli</i> BAP	0.6076	0.8333	0.4467	0.7667
<i>E. coli</i> MAC	0.9333	0.9333	1.0000	1.0000
<i>E. coli</i> MHA	0.9333	0.8667	1.0000	0.9000
<i>E. coli</i> MHB	0.3200	0.4000	0.3000	0.4000
<i>E. coli</i> NA	0.0000	0.6000	0.0000	0.6000
<i>E. cloacae</i> BAP	0.2238	0.1905	0.1900	0.2400
<i>E. cloacae</i> MAC	0.5333	0.5833	0.4667	0.5667
<i>E. cloacae</i> MHA	0.5333	0.5667	0.5000	0.5333
<i>E. cloacae</i> MHB	0.6476	0.5810	0.6000	0.5333
<i>E. cloacae</i> NA	0.6667	0.5333	0.6667	0.5000

Class	F1-A	F1-B	Prec-A	Prec-B
<i>K. pneumoniae</i> BAP	0.5238	0.6667	0.4000	0.6000
<i>K. pneumoniae</i> MAC	0.2600	0.0800	0.2333	0.0667
<i>K. pneumoniae</i> MHA	0.4667	0.5667	0.4000	0.5000
<i>K. pneumoniae</i> MHB	0.1333	0.3000	0.2000	0.2667
<i>K. pneumoniae</i> NA	0.8333	0.7333	0.8000	0.7667
<i>P. aeruginosa</i> BAP	0.5714	0.6000	0.5000	0.6000
<i>P. aeruginosa</i> MAC	0.5714	0.5333	0.5000	0.4667
<i>P. aeruginosa</i> MHA	0.6476	0.5810	0.5967	0.5333
<i>P. aeruginosa</i> MHB	0.6476	0.6476	0.5967	0.6800
<i>P. aeruginosa</i> NA	0.3667	0.4000	0.4000	0.5000

E.: *Escherichia* / *Enterobacter*; K.: *Klebsiella*; P.: *Pseudomonas*; A: Exp A; B: Exp B

The strongest performing classes were *E. coli* on MAC (F1 = 0.9333 in both experiments) and *E. coli* on MHA (F1 = 0.9333 for Experiment A, 0.8667 for Experiment B). *K. pneumoniae* on NA also performed strongly in Experiment A (F1 = 0.8333). These results are clinically significant because MacConkey agar and Mueller-Hinton agar are among the most commonly used media in clinical microbiology practice for the initial isolation and susceptibility testing of urinary tract infection pathogens, which are predominantly Gram-negative organisms. The aggregate confusion matrices for Experiment A and Experiment B are shown in Figures 5 and 6 respectively, compiled by accumulating predictions across all five cross-validation folds. The diagonal elements of both matrices confirm the strong performance on *E. coli* MAC and MHA, which show predominantly correct classifications (dark blue diagonal cells), while the off-diagonal elements reveal the specific inter-class confusions discussed below.

The weakest performing classes included *E. coli* NA (F1 = 0.0000 in Experiment A, recovering to 0.6000 in Experiment B), *K. pneumoniae* MHB (F1 = 0.1333 and 0.3000), *K. pneumoniae* MAC (F1 = 0.2600 and 0.0800), and *E. cloacae* BAP (F1 = 0.2238 and 0.1905). As visible in Figure 5, the complete failure of Experiment A on *E. coli* NA is evidenced by a fully empty diagonal cell for that class, with all test images from this class systematically misclassified as *E. coli* BAP across folds, suggesting that the frozen ImageNet feature extractor was unable to learn any discriminating feature between these two medium appearances. In contrast, Figure 6 shows that fine-tuning in Experiment B substantially corrected this failure, with the *E. coli* NA diagonal cell showing correct predictions, indicating that domain

adaptation was necessary to capture the subtle morphological differences between NA and BAP smears. Both Figures 5 and 6 also reveal that *K. pneumoniae* MAC and *E. cloacae* BAP

remain problematic in both experiments, with scattered off-diagonal misclassifications visible in the lower-left region of both matrices.

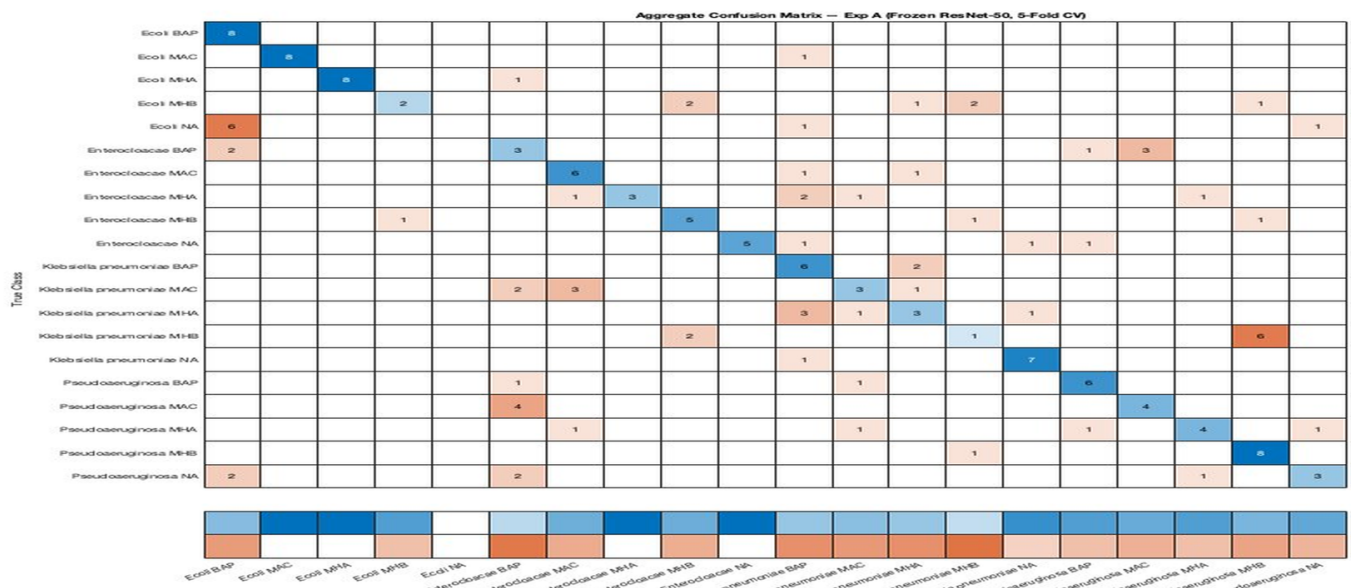


Figure 5. Aggregate Confusion Matrix — Experiment A (Frozen ResNet-50, 5-fold CV). Rows: True Class; Columns: Predicted Class.

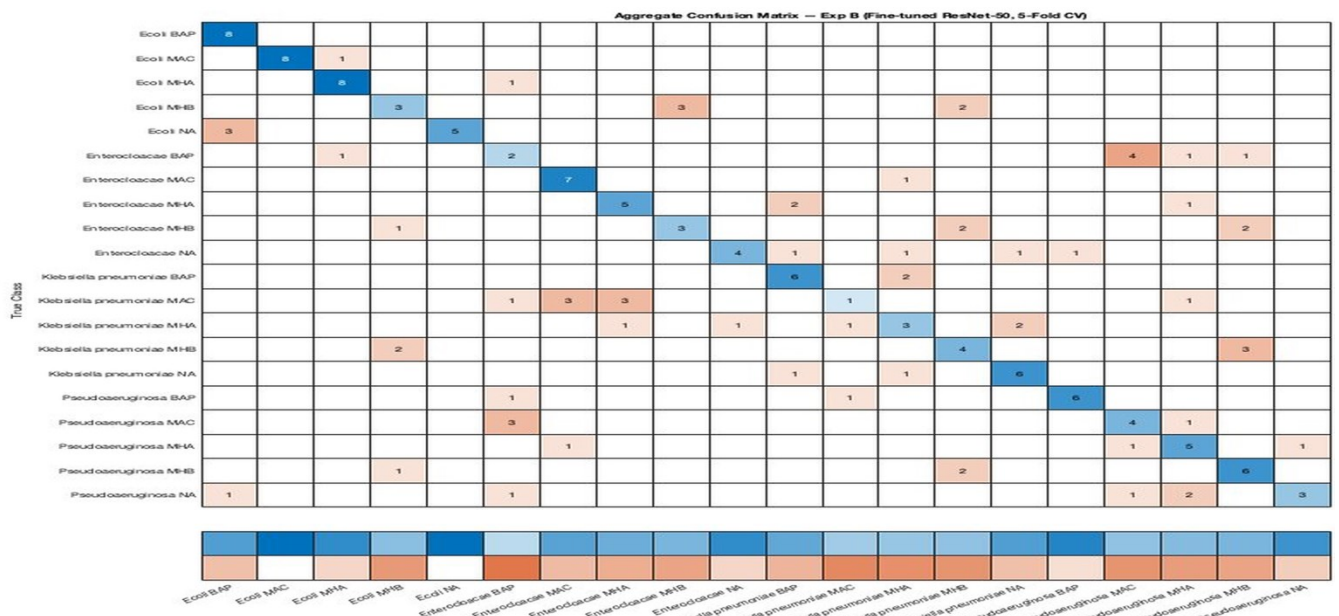


Figure 6. Aggregate Confusion Matrix — Experiment B (Fine-tuned ResNet-50, 5-fold CV). Rows: True Class; Columns: Predicted Class.

Beyond this descriptive pattern, the consistency of these specific failure modes across both the frozen and fine-tuned configurations is itself informative: because Experiment A and Experiment B share the same input images but differ substantially in trainable parameters, a misclassification that persists across both configurations is more likely to originate from a genuine ambiguity in the visual signal available to the model such as overlapping color and density characteristics

between *K. pneumoniae* and *E. cloacae* on MAC and BAP, both of which are also lactose-fermenting species with broadly similar bacilli morphology rather than from an architecture-specific limitation. This distinction matters because it suggests that simply increasing model capacity or fine-tuning depth, without addressing the underlying class overlap (for example through additional discriminative features, higher-resolution imaging, or a larger and more varied training set for these

specific classes), is unlikely to resolve these particular confusions in future iterations of this system.

C. Effect of Culture Medium Type on Classification Performance

The per-class results reveal a clear pattern: classification performance is strongly influenced by the visual characteristics of the Gram-stained smear produced by each culture medium, which varies independently of bacterial species. MacConkey agar (MAC) and Mueller-Hinton agar (MHA) consistently yielded the highest F1-scores, while Blood Agar Plate (BAP), Nutrient Agar (NA), and Mueller-Hinton Broth (MHB) produced lower and more variable performance. However, the reasons for strong performance differ between MAC and MHA and should be understood separately.

MacConkey agar is a selective and differential medium designed for Gram-negative enteric bacteria. It contains lactose, bile salts, crystal violet, and neutral red. Crystal violet and bile salts act as selective agents that inhibit the growth of Gram-positive organisms, while neutral red serves as a pH indicator that produces characteristic pink-to-red colony coloration in lactose-fermenting species such as *E. coli*, *E. cloacae*, and *K. pneumoniae*. Non-fermenting species such as *P. aeruginosa* produce colorless or pale colonies due to the absence of lactose fermentation. Although these differential colony characteristics are not reflected in Gram-stained smears, culture on MacConkey agar may influence the microscopic appearance of bacterial cells. Selective agents such as bile salts and crystal violet can affect bacterial growth characteristics and cellular morphology, potentially contributing to subtle interspecies differences observable on Gram staining. Such media-associated morphologic variation may partially explain the superior classification performance achieved using MAC-derived images. Combined with species-specific differences in cell morphology and arrangement visible under Gram staining, these characteristics provide strong visual discriminative features that ResNet-50 can exploit for accurate classification.

Mueller-Hinton agar, in contrast, is a non-selective general-purpose medium widely used in clinical microbiology primarily for antimicrobial susceptibility testing. It contains no inhibitory or differential agents and supports the growth of a broad range of organisms without producing color reactions. The reason MHA yields strong classification performance is therefore not selectivity. A plausible explanation is that the relatively standardized composition of MHA may promote more consistent bacterial growth conditions and physiological states across cultures. In contrast, enriched media such as BAP contain additional biological constituents that may influence bacterial growth dynamics and phenotypic expression. Such media-dependent differences may alter the microscopic appearance of bacteria in Gram-stained smears and consequently affect model discrimination performance. The superior performance observed with MHA-derived images may therefore reflect the influence of medium composition on bacterial phenotypic morphology and Gram-staining characteristics. However, the specific image features

responsible for this effect were not quantitatively analyzed in the present study and warrant further investigation.

Mueller-Hinton Broth, the only liquid medium evaluated, consistently produced the weakest results for *K. pneumoniae* and variable results for other species. Bacteria grown in liquid media are distributed throughout the broth in suspension rather than forming organized surface colonies on solid media. When smears are prepared from liquid cultures, the resulting bacterial distribution on the slide is typically more dispersed, less uniform, and produces more variable staining intensity compared to solid media cultures. This increased intra-class visual variability likely contributes to the lower classification performance observed for MHB classes.

D. Comparison of Transfer Learning Strategies

The overall superiority of fine-tuned ResNet-50 over frozen ResNet-50 in 5-fold cross-validation is consistent with transfer learning theory and existing empirical evidence [21]. When all layers are permitted to update, the network can progressively reshape its internal representations from those optimized for natural ImageNet photographs toward the texture patterns, staining characteristics, and morphological cues that distinguish bacterial species on Gram-stained slides at 1000× magnification. While ImageNet-pre-trained features include general-purpose edge, texture, and color detectors that partially generalize to microscopy images, the specific morphological features relevant for bacterial species discrimination (cell shape, arrangement, staining intensity distribution, background texture patterns) are not represented in ImageNet and require domain-specific adaptation.

However, the notably higher variance of Experiment B across folds reveals an important practical consideration. With approximately 113 training images across 20 classes (~5 per class) in each fold, fine-tuning all layers of a 25-million-parameter network creates a severely underdetermined optimization problem. The network has far more parameters than training examples, making it highly susceptible to overfitting the specific characteristics of the images present in each fold's training set. The frozen model, by preserving its pre-trained representations unchanged, avoids this problem at the cost of reduced domain adaptation. This variance-stability trade-off is a fundamental challenge in transfer learning for small medical datasets, and the empirical results presented here provide quantitative evidence for both sides of this trade-off in the specific context of Gram-stained bacterial image classification.

It is methodologically important to note that in a preliminary experiment using a single train/validation split without a separate held-out test set, the approach used in many published studies, Experiment A appeared to outperform Experiment B, with accuracy of 0.8000 versus 0.7500. This finding was completely reversed in 5-fold cross-validation, where Experiment B outperformed Experiment A by 0.0231 in accuracy. This reversal demonstrates that single-split evaluation is unreliable for comparing models on small datasets, where the outcome of a single split can be dominated by random variance in fold composition rather than true differences in model capability. The use of cross-validation is

therefore essential for valid model comparison in this data regime [22].

E. Statistical Significance and Baseline Model Comparison

To assess whether the performance difference between Experiment A and Experiment B is statistically meaningful rather than attributable to random fold-to-fold variation, a paired-samples t-test was conducted on the per-fold accuracy and macro F1-score values reported in Table 4 ($n = 5$ paired observations). The test yielded $t(4) = 0.49$, $p = 0.65$ for accuracy (mean difference = $+0.0232$, 95% CI $[-0.1085, 0.1549]$) and $t(4) = 0.93$, $p = 0.40$ for macro F1-score (mean difference = $+0.0399$, 95% CI $[-0.0786, 0.1584]$), indicating that neither difference is statistically significant at the conventional $\alpha = 0.05$ threshold. Because a paired t-test assumes normally distributed differences, an assumption that cannot be reliably verified with only five paired observations, a Wilcoxon signed-rank test was additionally conducted as a non-parametric alternative, yielding $p = 0.81$ for accuracy and $p = 0.63$ for F1-score, consistent with the t-test conclusion of no statistically significant difference. The corresponding paired effect sizes (Cohen's d) were 0.22 for accuracy and 0.42 for F1-score, both in the negligible-to-small range. Taken together, these results indicate that, given the present dataset size, the numerical advantage of fine-tuning over frozen feature extraction should be regarded as a small, statistically unconfirmed trend rather than a robust effect, reinforcing the need for validation on larger datasets before drawing firm conclusions about which strategy is preferable.

To evaluate whether ResNet-50 is a reasonable architectural choice for this task, a baseline comparison was conducted using SqueezeNet, a substantially more parameter-efficient CNN architecture (approximately 1.2 million parameters versus 25.6 million for ResNet-50), trained under identical 5-fold cross-validation conditions, data augmentation, and hyperparameters as Experiment B (Table 6). SqueezeNet (fine-tuned) achieved a mean accuracy of 0.4581 ± 0.1267 and macro F1-score of 0.3772 ± 0.1374 across the five folds, which is clearly lower than both ResNet-50 configurations reported in Table 3 (by 0.1020–0.1251 in accuracy and 0.1238–0.1637 in F1-score). This more pronounced gap, in contrast to the comparatively small and statistically non-significant difference observed between Experiment A and Experiment B, suggests that the residual connections and greater representational capacity of ResNet-50 provide a meaningful advantage over a substantially more compact architecture for this Gram-stained multi-medium classification task, and that the choice of ResNet-50 is not arbitrary. The wider per-fold variability of SqueezeNet (accuracy std = 0.1267, the highest of all three configurations evaluated) further suggests that lower-capacity architectures may be especially sensitive to the limited training data available per fold in this study. A more exhaustive architecture search, including additional backbones such as DenseNet, EfficientNet, and Vision Transformers, is left for future work with a larger dataset where such comparisons can be made with greater statistical confidence.

Table 6. Mean $\pm$ Standard Deviation Comparison Across Three Model Configurations (5-Fold CV)

Configuration	Accuracy	Macro F1-Score
Exp A: ResNet-50 (Frozen)	0.5601 ± 0.0603	0.5010 ± 0.0935
Exp B: ResNet-50 (Fine-tuned)	0.5832 ± 0.1109	0.5409 ± 0.1231
SqueezeNet (Fine-tuned)	0.4581 ± 0.1267	0.3772 ± 0.1374

IV. CONCLUSION

This study presented a comparative evaluation of frozen and fine-tuned ResNet-50 transfer learning for the automated classification of Gram-stained bacterial microscopy images across four clinically significant species (*Enterobacter cloacae*, *Escherichia coli*, *Klebsiella pneumoniae*, and *Pseudomonas aeruginosa*) and five culture media types (BAP, MAC, MHA, MHB, NA), formulated as a 20-class problem and rigorously evaluated using stratified 5-fold cross-validation on a primary dataset of 166 images.

Fine-tuned ResNet-50 (Experiment B) outperformed frozen feature extraction (Experiment A) overall, achieving mean accuracy of 0.5832 ± 0.1109 and macro F1-score of 0.5409 ± 0.1231 versus 0.5601 ± 0.0603 and 0.5010 ± 0.0935 , although a paired-samples statistical test indicated that this difference is not statistically significant given the small number of cross-validation folds. The highest per-class F1-scores of 0.9333 were achieved for *E. coli* on MAC and MHA media in both experiments. Culture medium type was identified as the primary determinant of per-class classification difficulty: selective and differential media (MAC, MHA) produced the strongest results, while non-selective (BAP, NA) and liquid (MHB) media produced weaker and more variable performance. It must be explicitly acknowledged that the primary dataset used in this study, comprising only 166 images across 20 classes (8–9 images per class, and approximately 5–6 images per class available for training within each cross-validation fold), is small relative to the scale typically required for robust deep learning generalization. The moderate overall accuracy (approximately 58%) obtained despite the use of online data augmentation indicates that the trained models have not yet reached a level of generalization sufficient for unsupervised clinical deployment, and the present results should be regarded as a methodological baseline rather than a clinically deployable solution.

Three principal contributions are reported: (1) the first systematic evaluation of ResNet-50 transfer learning across five culture media types for Gram-negative bacterial species using Gram-stained microscopy images; (2) the finding that culture medium selectivity and differential properties are key determinants of automated classification performance, with direct implications for clinical workflow design; and (3) the methodological demonstration that cross-validated evaluation is essential for reliable model comparison on small datasets,

as single-split evaluation produced rankings that were reversed by 5-fold cross-validation.

Acquiring a substantially larger volume of raw, real-world Gram-stained microscopy images remains the most critical priority for future work: the present results indicate that data augmentation alone cannot fully compensate for the limited number of primary images per class, and future work should expand the dataset to at least 50–100 real (non-augmented) images per class through additional clinical specimen collection at multiple laboratory sites. Equally important, the explainability limitation noted in Section III.F should be directly addressed: given that overall accuracy remains moderate (approximately 58%), the application of Grad-CAM visualization [28] or comparable attention-mapping techniques is not merely a desirable enhancement but a necessary diagnostic step to verify whether the trained models are genuinely learning discriminative bacterial morphological features or are instead partly relying on incidental background characteristics of the culture medium; without this verification, confidence in the clinical relevance of the reported accuracy figures remains limited. Beyond these two priorities, additional network architectures including EfficientNet, DenseNet, and Vision Transformers should be more exhaustively evaluated with statistical significance testing once a larger dataset is available, and multi-site validation with images from different microscopes, operators, and staining batches will be necessary to assess generalizability for real-world clinical deployment.

REFERENCES

- [1] World Health Organization, "Global antimicrobial resistance and use surveillance system (GLASS) report 2022," WHO, Geneva, Switzerland, 2022.
- [2] H. C. Gram, "Ueber die isolirte Färbung der Schizomyceten," *Fortschritte der Medicin*, vol. 2, pp. 185–189, 1884.
- [3] C. A. Petti, C. R. Polage, T. C. Quinn, A. R. Ronald, and M. A. Sande, "Laboratory medicine in Africa: a barrier to effective health care," *Clin. Infect. Dis.*, vol. 42, no. 3, pp. 377–382, 2006.
- [4] M. Z. Alom et al., "A state-of-the-art survey on deep learning theory and architectures," *Electronics*, vol. 8, no. 3, p. 292, 2019.
- [5] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, 2010.
- [6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
- [7] A. Mairi, L. Hamza, and A. Touati, "Artificial intelligence and its application in clinical microbiology," *Expert Rev. Anti-infect. Ther.*, vol. 23, no. 7, pp. 469–490, Mar. 2025, doi: 10.1080/14787210.2025.2484284.
- [8] K. P. Smith and J. E. Kirby, "Image analysis and artificial intelligence in infectious disease diagnostics," *Clin. Microbiol. Infect.*, vol. 26, no. 10, pp. 1318–1323, 2020.
- [9] J. Dien Bard et al., "Proceedings of the Clinical Microbiology Open 2024: Artificial intelligence applications in clinical microbiology," *J. Clin. Microbiol.*, 2025, doi: 10.1128/jcm.01965-24.
- [10] B. Zielinski et al., "Deep learning approach to bacterial colony classification," *PLOS ONE*, vol. 12, no. 9, p. e0184554, 2017.
- [11] S. Visitsattapong, M. Bunkum, C. Pintavirooj, and M. P. Paing, "A deep learning model for bacterial classification using Big Transfer (BiT)," *IEEE Access*, vol. 12, pp. 15609–15621, 2024, doi: 10.1109/ACCESS.2024.3358671.
- [12] A. J. Abougarair, A. Oun, S. I. Sawan, and A. Ma'arif, "Deep learning-based automated approach for classifying bacterial images," *Int. J. Robot. Control Syst.*, vol. 4, no. 2, pp. 849–876, Jun. 2024.
- [13] E. Gullu, S. Bora, and B. Beynek, "Exploiting image processing and artificial intelligence techniques for the determination of antimicrobial susceptibility," *Appl. Sci.*, vol. 14, no. 9, 2024, doi: 10.3390/app14093950.
- [14] V. Shwetha et al., "Vision transformer based bacteria classification model for Gram-stained direct smear images," *Multimed. Tools Appl.*, vol. 84, pp. 20289–20309, 2025, doi: 10.1007/s11042-025-20742-0.
- [15] M.-H. Lee, Y.-T. Tsai, T.-Y. Huang, D.-J. Liew, I. Candia, and Y.-C. Chang, "Alignment-free bacterial pathogen identification using vision transformer and image augmentation techniques in high-resolution microscopy," *Chinese J. Chem. Eng.*, vol. 88, pp. 142–151, 2025.
- [16] S. Y. Chin, J. Dong, K. Hasikin, R. Ngui, K. W. Lai, P. S. Q. Yeoh, and X. Wu, "Bacterial image analysis using multi-task deep learning approaches for clinical microscopy," *PeerJ Comput. Sci.*, vol. 10, p. e2180, Aug. 2024, doi: 10.7717/peerj-cs.2180.
- [17] M. Talo, "Automated classification of histopathology images using transfer learning," *Artif. Intell. Med.*, vol. 101, p. 101743, 2019.
- [18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015.
- [19] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.
- [20] N. Bayramoglu and J. Heikkilä, "Transfer learning for cell nuclei classification in histopathology images," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2016, pp. 532–539.
- [21] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" *Adv. Neural Inf. Process. Syst.*, vol. 27, pp. 3320–3328, 2014.
- [22] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc.*



- Int. Joint Conf. Artif. Intell.*, vol. 14, no. 2, 1995, pp. 1137–1145.
- [23] P. R. Murray, K. S. Rosenthal, and M. A. Pfaller, *Medical Microbiology*, 9th ed. Amsterdam, Netherlands: Elsevier, 2021.
- [24] A. Telea, "An image inpainting technique based on the fast marching method," *J. Graph. Tools*, vol. 9, no. 1, pp. 23–34, 2004.
- [25] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
- [26] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.
- [27] A. Chaddad et al., "Generalizable and explainable deep learning for medical image computing: An overview," *Curr. Opin. Biomed. Eng.*, vol. 34, p. 100581, 2025.
- [28] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.