TELKOMNIKA Telecommunication Computing Electronics and Control
Vol. 24, No. 6, December 2026, pp. 2012~2022
ISSN: 1693-6930, DOI: 10.12928/TELKOMNIKA.v24i6.28022

HieroBact: clinically coherent hierarchical bacterial classification from Gram-stained images

Sepyan Purnama Kristanto[1], Moh. Erdda Habiby[2], Lutfi Hakim[1], Dianni Yusuf[1]
[1]Department of Information Technology, Politeknik Negeri Banyuwangi, Banyuwangi, Indonesia
[2]School of Computer Science, Universitas Bina Nusantara (BINUS), Jakarta, Indonesia

Article Info

Article history:

Received May 11, 2026
Revised Aug 4, 2026
Accepted Aug 28, 2026

Keywords:

Bacterial image classification
Gram-stained microscopy
Hierarchical multi-task learning
Hierarchically conditioned loss
Taxonomic coherence

ABSTRACT

Gram-stained microscopy provides rapid preliminary evidence for clinical microbiology, but most deep learning systems still classify bacterial species as a flat label set and therefore ignore the Gram-reaction and morphology hierarchy that guides early antimicrobial decisions. This paper presents HieroBact, a hierarchical multi-task learning (MTL) framework that jointly predicts Gram reaction, cellular morphology, and species identity using DIBaS-Tax annotations, an image-only feature-fusion backbone, and a hierarchically conditioned loss (HC-Loss) that reduces species-level supervision when coarse-level confidence is unreliable. Under 5-fold cross-validation on DIBaS-Tax, HieroBact achieves 99.1% Gram-level, 97.8% morphology-level, and 97.2% species-level accuracy, while reducing clinically unsafe Gram-level misclassifications from 28.8% to 5.7% compared with a strong flat baseline. The results show that explicitly modelling bacterial taxonomy improves not only recognition accuracy but also the clinical coherence and interpretability of model errors, making HieroBact a practical step toward safer diagnostic-support systems for preliminary laboratory decision making and subsequent confirmation by trained clinical microbiologists.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Sepyan Purnama Kristanto
Department of Information Technology, Politeknik Negeri Banyuwangi
Raya Jember St. Km 13, Banyuwangi 68461, East Java, Indonesia
Email: sepyan@poliwangi.ac.id

1. INTRODUCTION

Bacterial infectious diseases remain a major global health burden, responsible for an estimated 1.5 million Diarrhoeal deaths annually, predominantly in resource-limited settings [1]. Rapid identification of causative organisms is therefore central to timely antimicrobial therapy. Gram staining is one of the most widely used preliminary diagnostic procedures in clinical microbiology, producing actionable information within 5–10 minutes of specimen preparation [2]. It separates bacteria into Gram-positive (G+, crystal violet retained) and Gram-negative (G-, safranin counterstained) groups and reveals cellular morphology, such as rod, coccus, vibrio, or filamentous forms [3]. Because Gram reaction and morphology inform empirical antibiotic selection, a Gram-level classification error is not merely taxonomic; it may direct treatment toward an inappropriate antimicrobial class [4]. This clinical risk necessitates automated systems whose errors remain biologically and clinically coherent.

Deep learning has achieved strong performance on Gram-stained bacterial image benchmarks, but most reported systems remain accuracy-centered. On digital image of bacterial species (DIBaS) [5], transfer learning and feature-fusion backbones have produced high species-level accuracy [6], [7]. Subsequent studies have examined feature-selection transfer learning, big transfer, compact depth wise-separable convolutional

neural networks (CNNs), capsule networks, and vision transformers [8]–[12]. Although these methods differ in backbone and training strategy, they generally treat bacterial identification as a flat species classification problem. Consequently, a prediction can be scored as wrong without distinguishing a within-Gram species error from a clinically more serious cross-Gram error.

Structure-aware medical-image classification shows that explicit label dependencies can improve the consistency of model errors [13]. Recent bacterial recognition studies have also broadened the evidence base through phase-contrast, fluorescence, multi-media Gram-stained, hyperspectral, Raman, and surface-enhanced Raman spectroscopy (SERS) imaging [14]–[22]. In parallel, hierarchical and consistency-aware models [23], [24] and explainable artificial intelligence (XAI) in medical imaging [25], [26] emphasize that clinical models should expose meaningful intermediate decisions rather than only a terminal class label. Bacteriological references such as Bergey's manual provide the relevant coarse-to-fine taxonomy [27], while multi-task and adaptive loss-weighting methods provide mechanisms for shared representation learning [28]–[30]; curriculum learning further motivates coarse-to-fine supervision [31]. However, these lines of research have not been integrated into a bacterial microscopy classifier that jointly predicts Gram reaction, morphology, and species while directly measuring the clinical coherence of residual errors.

The research gap addressed in this study is therefore the absence of a bacterial microscopy classifier that jointly predicts species labels and preserves the clinically meaningful Gram-reaction and morphology hierarchy. This study aims to build and evaluate such a hierarchy-aware classifier for Gram-stained bacterial images. The contributions are fourfold: (i) HieroBact architecture: a three-head hierarchical multi-task network using an image-only AquaFusionNet-inspired feature-fusion backbone [7], predicting Gram reaction, morphology, and species identity in a single forward pass; (ii) hierarchically conditioned loss (HC-Loss): a confidence-conditioned training objective that scales species-level supervision by the model's current Gram and morphology confidence, yielding a self-regulated coarse-to-fine learning process; (iii) DIBaS-Tax: Gram-reaction and cellular-morphology annotations for all 33 DIBaS species, cross-verified against independent microbiology references, and (iv) taxonomic coherence metric (TCM) and morphological coherence metric (MCM): two coherence metrics that quantify whether species errors remain within the correct Gram group or morphology class, complementing conventional accuracy-based evaluation.

2. THE PROPOSED HIEROBACT FRAMEWORK

2.1. Dataset and DIBaS-Tax annotation

The DIBaS dataset [5] comprises 660 RGB Gram-stained micrographs from 33 bacterial species (≈20 images/species), captured at 100× oil-immersion magnification under a standardised staining protocol. The dataset is class-balanced, publicly available, and covers clinically relevant species from waterborne, respiratory, urogenital, and wound-infection contexts. Figure 1 shows representative examples arranged by Gram reaction and cellular morphology. Rows: Gram-positive (G+, purple border, crystal violet) and Gram-negative (G-, red border, safranin). Columns: rod, cocci, vibrio, filamentous/spirilla. G+ vibrio does not occur in DIBaS.

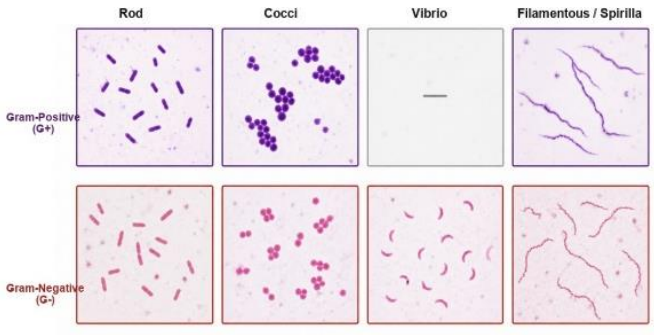


Figure 1. Representative Gram-stained microscopy images from DIBaS-Tax

DIBaS-Tax augmented every DIBaS image with two additional species-level labels: Gram reaction ($y_g \in$ {G+, G-}) and cellular morphology ($y_m \in$ {rod, cocci, vibrio, filamentous/spirilla}). Labels were assigned at the species level and cross-verified against Murray *et al.* [3] and Bergey's manual [27]. Agreement between these two sources was 100% across all 33 species. Four species with historically contested staining profiles (*Mycobacterium smegmatis*, *Gardnerella vaginalis*, *Propionibacterium acnes*,

Actinomyces israelii) were additionally checked by a practising clinical microbiologist. This process ensured annotation reliability while preserving the standard taxonomic assumption that all images of a species share identical Gram and morphology labels. Table 1 summarises the resulting label distribution, and Figure 2 presents the three-level taxonomy tree used for training and evaluation.

Pre-processing: images were resized to 224×224 px and normalised with ImageNet channel statistics ($\mu = [0.485,0.456,0.406]$, $\sigma = [0.229,0.224,0.225]$). Training folds used random horizontal/vertical flips ($p = 0.5$), rotations (±15°), colour jitter (brightness/contrast ±20%, hue ±5%), and Gaussian blur ($\sigma \sim U[0.1,0.5]$). Hyperparameters were selected from a small validation sweep guided by species accuracy and taxonomic coherence; the final configuration used the best accuracy–coherence trade-off in the neighbouring-weight sensitivity analysis. Evaluation used 5-fold stratified cross-validation, and all reported metrics are mean ± standard deviation across folds.

Table 1. DIBaS-Tax label distribution across three taxonomic levels

Level	Category	Species	Images (%)
Gram reaction	G+ (Gram-positive)	20	400 (60.6%)
	G- (Gram-negative)	13	260 (39.4%)
Morphology	Rod	22	440 (66.7%)
	Cocci	9	180 (27.3%)
	Vibrio	1	20 (3.0%)
	Filamentous/spirilla	1	20 (3.0%)
Species	(Total)	33	660 (100%)

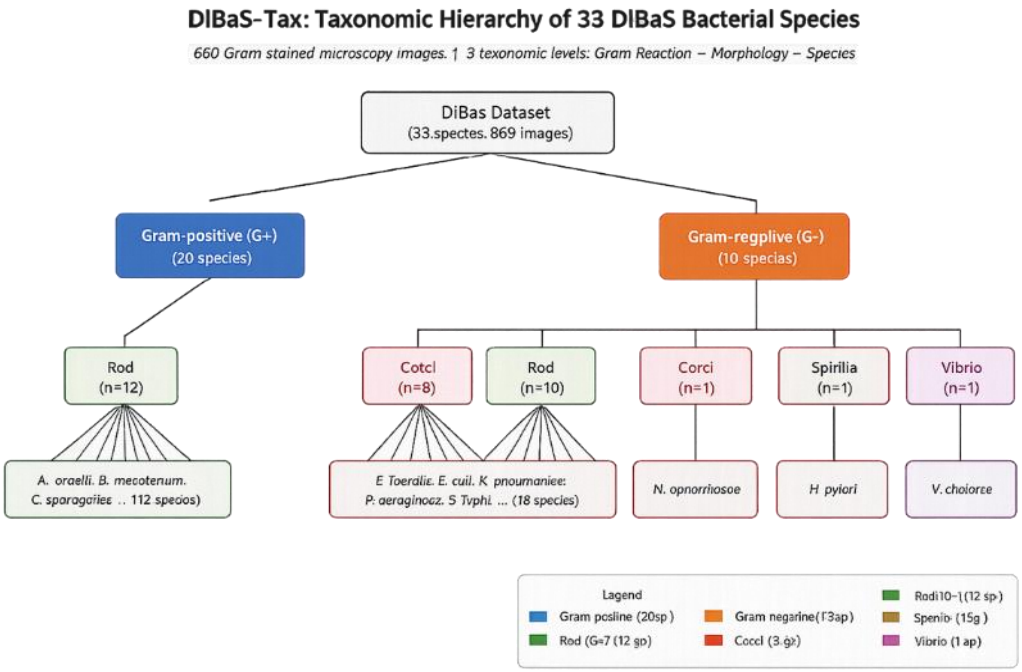


Figure 2. DIBaS-Tax three-level taxonomy tree; level 1: Gram reaction (G+: 20 sp.; G-: 13 sp.), level 2: cellular morphology; and leaves: individual species with image count

2.2. HieroBact architecture

HieroBact is built around a shared feature extractor that feeds three parallel classification heads, one per taxonomic level. Figure 3 summarises both the architecture and training workflow: panel (a) shows the augmented microscopy input, panels (b)–(d) show shared feature extraction and fusion, and panels (e)–(f) show task-specific outputs, HC-Loss supervision, and the inference-time consistency mask. During each training step, one shared representation is passed to all heads, the three cross-entropy losses are computed, HC-Loss conditions the species term on detached coarse-level confidence, and Adam with decoupled weight decay (AdamW) updates the shared extractor and all task-specific branches end-to-end.

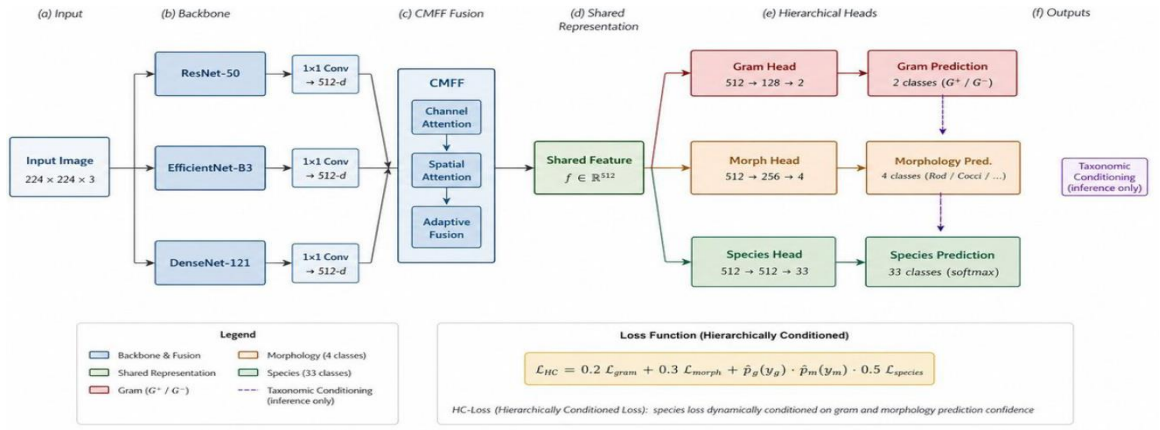


Figure 3. HieroBact architecture and training workflow

Shared feature extractor: the backbone implements an image-only adaptation of AquaFusionNet-style feature fusion [7]. Three CNN backbones (ResNet-50 [32], EfficientNet-B3 [33], DenseNet-121 [34]) each project their spatial feature maps to 512 channels via 1×1 convolution. The cross-modal feature fusion (CMFF) module then combines the projected maps via sequential channel attention, spatial attention, and adaptive weighted summation:

$$f = \mathrm{GAP}\left(\sum_{k=1}^{3} \alpha_k \cdot w_s^k \odot w_c^k \odot \tilde{F}_k\right), \quad f \in \mathbb{R}^{512} \tag{1}$$

where w_c^k and w_s^k are channel and spatial attention weights, α_k are softmax-normalised backbone fusion weights, and $GAP(\cdot)$ is global average pooling. In computational terms, CMFF adds lightweight 1×1 projections and attention operations after the three backbone feature maps have been extracted; therefore, the main inference cost remains dominated by the backbone ensemble rather than by the fusion block itself. This design keeps the hierarchy-aware components modular: the shared feature extractor can be replaced by a lighter CNN if deployment constraints become more important than peak species accuracy.

Hierarchical heads: three fully-connected heads with rectified linear unit (ReLU) activations and SoftMax outputs operate on the shared representation f:

– Gram head h_g: FC (512→128) – ReLU – Dropout (0.3) – FC (128→2)
– Morphology head h_m: FC (512→256) – ReLU – Dropout (0.3) – FC (256→4)
– Species head h_s: FC (512→512) – ReLU – Dropout (0.5) – FC (512→33)

The heads learn complementary outputs from the same fused image representation, so coarse and fine predictions are available in one forward pass. Their separate parameters allow each taxonomic level to retain an appropriate output capacity.

Inference-time taxonomic mask: at inference, species probabilities inconsistent with the argmax Gram and morphology predictions are zeroed and the distribution re-normalised:

$$\hat{p}_s{}'(c) = \frac{\hat{p}_s(c)\,1[c \in \mathcal{C}(\hat{y}_g, \hat{y}_m)]}{\sum_{c'} \hat{p}_s(c')\,1[c' \in \mathcal{C}(\hat{y}_g, \hat{y}_m)]}$$
$$\hat{p}_s{}'(c) = \frac{\hat{p}_s(c)\,1[c \in \mathcal{C}(\hat{y}_g, \hat{y}_m)]}{\sum_{c'} \hat{p}_s(c')\,1[c' \in \mathcal{C}(\hat{y}_g, \hat{y}_m)]} \tag{2}$$

where $\mathcal{C}(\hat{y}_g, \hat{y}_m)$ is the set of species consistent with $\hat{y}_g$ and $\hat{y}_m$. This mask is applied only at inference and imposes zero training overhead. It therefore improves cross-level consistency without altering the learned parameters.

2.3. HC-Loss

Standard fixed-weight multi-task learning (MTL) combines task losses as:

$$\mathcal{L}_{\mathrm{H}} = \lambda_1 \mathcal{L}_g + \lambda_2 \mathcal{L}_m + \gamma(t) \mathcal{L}_s$$
$$\gamma(t) = \lambda_3 \cdot \hat{p}_g(y_g) \cdot \hat{p}_m(y_m), \quad \lambda_1 = 0.2,\ \lambda_2 = 0.3,\ \lambda_3 = 0.54 \tag{3}$$

where $\mathcal{L}_g$, $\mathcal{L}_m$, $\mathcal{L}_s$ are cross-entropy losses for the Gram, morphology, and species heads.

A static weighting scheme in (3) provides constant species-level gradients regardless of coarse-level resolution. Asking a network to separate 33 species before it can reliably distinguish G+ from G- imposes fine-grained supervision before the coarse representation is stable. When Gram accuracy hovers near 50%, species-level gradients carry little useful signal and risk pulling shared representations in contradictory directions. HC-Loss replaces the fixed species coefficient with a dynamic scalar:

The base weights ($\lambda_1 = 0.2$, $\lambda_2 = 0.3$, $\lambda_3 = 0.5$) assign the largest fixed coefficient to species (primary objective) while allocating smaller weights to the faster-converging auxiliary tasks. Their sum $\sum\lambda_i = 1$ bounds the total gradient magnitude regardless of $\gamma(t)$. The selected setting was not treated as an arbitrary constant: a neighbouring-weight sensitivity analysis evaluates neighbouring configurations and shows that species accuracy changes by only 0.2 pp, whereas coherence metrics provide the clearer criterion for choosing the final trade-off, where y_g, y_m are the ground-truth Gram and morphology labels, and $\hat{p}_g(y_g)$, $\hat{p}_m(y_m)$ are the model's current predicted probabilities for those correct labels. At training start with chance-level coarse accuracy, $\gamma(t) \approx 0.5 \times 0.5 \times 0.25 = 0.0625$, strongly suppressing species gradients. As coarse accuracy improves, $\gamma(t)$ grows monotonically toward $\lambda_3 = 0.5$, progressively restoring full species-level supervision. The conditioning scalars $\hat{p}_g(y_g)$ and $\hat{p}_m(y_m)$ are detached from the computation graph $via.detach()$: this ensures that backpropagated gradients through $\mathcal{L}_s$ are not erroneously attributed to the Gram and morphology heads, preserving clean gradient attribution between heads.

2.4. TCM

Standard species accuracy does not distinguish between errors that are clinically benign (correct Gram group, wrong species) and those that are clinically dangerous (wrong Gram group). Two complementary metrics are defined. Both are calculated only over incorrectly classified species samples, thereby separating error coherence from overall accuracy.

The TCM is the fraction of incorrectly classified species samples $\mathcal{E}$ whose predicted species belongs to the correct Gram group:

$$\mathrm{TCM} = \frac{|\{(x,y)\in\mathcal{E}:\mathrm{gram}(\hat{y}_s)=\mathrm{gram}(y_s)\}|}{|\mathcal{E}|}$$
$$\mathrm{TCM} = \frac{|\{(x,y)\in\mathcal{E}:\mathrm{gram}(\hat{y}_s)=\mathrm{gram}(y_s)\}|}{|\mathcal{E}|} \quad (4)$$

The MCM analogously measures the fraction of species errors within the correct morphology class:

$$\mathrm{MCM} = \frac{|\{(x,y)\in\mathcal{E}:\mathrm{morph}(\hat{y}_s)=\mathrm{morph}(y_s)\}|}{|\mathcal{E}|}$$
$$\mathrm{MCM} = \frac{|\{(x,y)\in\mathcal{E}:\mathrm{morph}(\hat{y}_s)=\mathrm{morph}(y_s)\}|}{|\mathcal{E}|} \quad (5)$$

Both metrics range in $[0,1]$; higher is better. A system with TCM=1.0 never commits a Gram-incoherent error regardless of species accuracy. Metric validity was assessed by applying TCM and MCM to the flat baseline, fixed-weight MTL, and the proposed HC-Loss model. The metrics increased when the model architecture explicitly constrained errors to remain within biologically consistent groups, supporting their use as complementary clinical-safety indicators rather than replacements for species accuracy.

2.5. Implementation details

All experiments used PyTorch 2.1.0 (CUDA 12.1). Backbone weights were initialised from ImageNet-21k checkpoints [35], and all branches were fine-tuned end-to-end. The optimiser was AdamW [36] ($\eta_0 = 10^{-4}$, weight decay 5×10^{-5}) with cosine annealing and warm restarts [37] ($T_0 = 10$, $T_{mult} = 2$). Batch size was set to 32, the maximum training length was 100 epochs, and early stopping used patience 15 monitored on validation species accuracy. Loss weights were selected through the neighbouring-weight sensitivity analysis rather than tuned on the test folds. Experiments ran on an NVIDIA RTX 3090 (24 GB VRAM), with a mean training time of 4.2 h/fold. The full three-backbone model contains approximately 47.2M parameters; this is feasible for workstation or hospital-laboratory GPU deployment, but it is not yet an embedded edge model. Statistical significance was tested using the paired two-tailed McNemar test [38] ($\alpha = 0.05$, Bonferroni-corrected for three pairwise comparisons).

Together, the annotation protocol, hierarchical heads, HC-Loss, and coherence metrics form a reproducible coarse-to-fine evaluation pipeline. The design separates training-time confidence conditioning from inference-time taxonomic masking so that their contributions can be evaluated independently. The next section reports classification, coherence, ablation, and robustness results.

3. RESULTS AND DISCUSSION

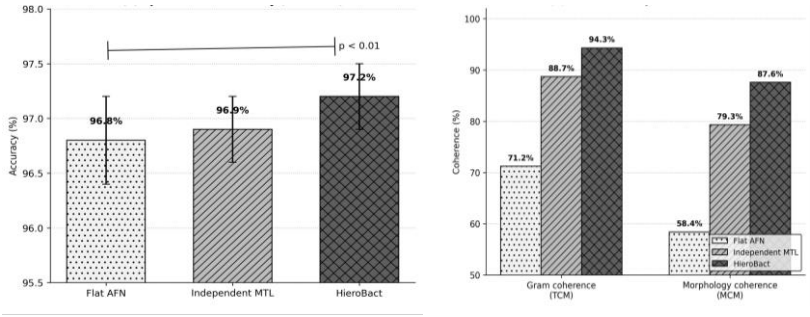
3.1. Per-level classification performance

Table 2 reports accuracy at all three taxonomic levels together with species-level macro-F1 and TCM under 5-fold cross-validation; MCM is summarized in Figures 4(a) and (b). The flat baseline provides species predictions only, whereas the two multi-task models also report Gram and morphology outputs. All values were aggregated across identical stratified folds. Class-specific sensitivity and specificity are not reported because complete out-of-fold prediction vectors were not retained. Table 2. Per-level classification performance (5-fold CV, mean $\pm$ std). Flat AFN = image-only AquaFusionNet-style feature-fusion baseline [7] without hierarchical heads. Indep. MTL = same three-head architecture with fixed weights ($\lambda_1 = 0.2$, $\lambda_2 = 0.3$, $\lambda_3 = 0.5$) throughout training, no dynamic conditioning. $*p < 0.05$, $**p < 0.01$ vs. Flat AFN (McNemar, Bonferroni-corrected).

HieroBact achieves near-ceiling Gram-level performance (99.1%), consistent with the high visual discriminability of Gram staining under standardised laboratory conditions. The 0.4 pp gain in species-level identification over Flat AFN is modest in absolute terms but statistically significant ($p < 0.01$, McNemar); the 0.3 pp improvement over Indep. MTL is significant at $p < 0.05$, confirming that dynamic HC-Loss conditioning provides a genuine, if modest, benefit beyond fixed-weight hierarchical supervision. More importantly, the flat baseline is already a strong image-only AquaFusionNet-style model [7] with ImageNet-21k initialisation, making any improvement non-trivial. The primary benefit of HieroBact lies not in raw accuracy but in the clinical quality of its errors, examined next. Figure 5 presents the Gram-reaction matrix reconstructed from retained aggregate counts. DIBaS-Tax contains 400 G+ and 260 G- images; six errors yield 654/660 correct predictions (99.1%). Morphology- and species-level matrices are omitted because complete out-of-fold predictions were not retained.

Table 2. Per-level classification performance

Method	Gram (%)	Morph. (%)	Sp. (%)	F1	TCM	Params
Flat AFN [7]	—	—	96.8±0.4	0.967	71.2	45.8 M
Indep. MTL	98.5±0.3	97.1±0.4	96.9±0.3*	0.968	88.7	47.2 M
HieroBact	99.1±0.2	97.8±0.3	97.2±0.3**	0.971	94.3	47.2 M



(a) (b)

Figure 4. HieroBact vs. baselines (5-fold CV, mean $\pm$ 1 SD): (a) species accuracy; significance markers from McNemar test and (b) TCM (+23.1 pp) and MCM (+29.2 pp) over the flat baseline. Higher coherence indicates more clinically appropriate error patterns

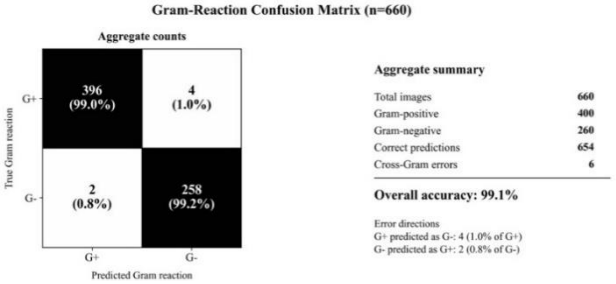


Figure 5. HieroBact Gram-reaction confusion matrix for 660 DIBaS-Tax images. Rows sum to 400 G+ and 260 G- images; 654 predictions are correct (99.1%)

3.2. Taxonomic coherence

TCM and MCM are the metrics that most directly capture clinical utility. TCM evaluates whether a species error remains within the correct Gram group, while MCM evaluates whether it remains within the correct morphology class. Figure 4 summarizes both coherence measures alongside species accuracy, highlighting that the larger gain lies in error coherence.

One in every 3.5 misclassifications by the flat baseline crosses the Gram-positive/negative boundary. HieroBact compresses that proportion from 28.8% to 5.7%, an 80% relative reduction. In practical per-fold terms (≈132 test images), this corresponds to 3–4 Gram-incoherent errors for HieroBact versus 12–14 for the flat baseline. The MCM improvement (58.4%→87.6%) shows that when HieroBact errs on species, the predicted species almost always shares the correct morphological class, a failure mode carrying substantially lower clinical consequence. Figure 4 compares this pattern visually: Figure 4(a) shows the species-accuracy margin, whereas Figure 4(b) shows that the larger gain comes from improved taxonomic and morphological coherence.

3.3. Ablation study

Table 3 isolates the contribution of each component. Each row adds one mechanism to the preceding configuration while preserving the shared feature extractor. The comparison therefore distinguishes the effects of hierarchical heads, HC-Loss, and inference-time masking.

The most impactful step is adding hierarchical heads (A→D): TCM improves by 18.9 pp with only +0.2 pp species accuracy gain, confirming that hierarchical supervision primarily restructures the distribution of misclassifications rather than boosting accuracy. HC-Loss (D→E) further adds $+2.7$ pp TCM and $+0.1$ pp accuracy, demonstrating that confidence-based dynamic weighting provides an additional benefit beyond fixed-weight MTL. The inference-time mask (E→F) contributes +1.5 pp TCM at zero training overhead. Neighboring HC-Loss weights changed species accuracy by at most 0.2 percentage points. The selected weights (0.2, 0.3, 0.5) gave the best accuracy-coherence trade-off, indicating that the result was not driven by a single unstable setting.

Table 3. Step-wise ablation. Each row adds one component to the previous

Configuration	Sp. Acc.	TCM
Flat AFN (baseline)	96.8	71.2%
+ Gram head (fixed λ)	96.9	89.4%
+ Morph head (fixed λ)	96.9	88.7%
All 3 heads, fixed λ	97.0	90.1%
+ HC-Loss, no mask	97.1	92.8%
+ Taxonomic mask (full model)	97.2	94.3%

3.4. Comparison with state-of-the-art

Table 4 places HieroBact in context. Several verified DIBaS studies report higher flat species accuracy than HieroBact, but they do not provide coarse-level Gram and morphology outputs, do not evaluate taxonomic coherence, and therefore cannot directly support empirical antibiotic decision making. HieroBact is the only system in this comparison that jointly evaluates species accuracy, robustness, and taxonomic coherence.

Recent journal studies also show that bacterial recognition is moving beyond single benchmark classification. Label-free phase-contrast microscopy, fluorescence microscopy for tuberculosis, multi-media Gram-stained microscopy, and YOLOv8-based Gram-stained detection all demonstrate the growing interest in automated bacterial image analysis [26]–[29]. Spectral and hyperspectral approaches further show that deep learning can identify pathogenic bacteria under different sensing modalities, including single-cell Raman spectra, hyperspectral microscopy, unsupervised Raman clustering, SERS-based pathogen identification, and clinical Raman diagnostics [30]–[34]. These studies strengthen the broader motivation for automated microbiology, but they still rarely evaluate whether model errors preserve clinically meaningful Gram and morphology relationships.

The closest related approaches are hierarchical classifiers and robustness methods for medical images. Recent work on hierarchical chest X-ray taxonomies and consistency-aware hierarchical image classification supports the value of structured labels [23], [24], while robustness studies emphasize that clinical models should be tested under adversarial or real degradation shifts [39], [40]. Multi-task bronchoscopy image classification illustrates that shared visual representations can support multiple clinical outputs [41]. In bacterial identification, generative adversarial network (GAN)-assisted Raman classification and laser-tweezers Raman deep learning provide additional evidence that robustness and generalisation

remain central challenges in microbial recognition [42], [43]. HieroBact differs from these directions by placing the microbiological hierarchy itself inside both training and evaluation.

Table 4. State-of-the-art comparison on DIBaS-oriented bacterial classification

Method	Backbone	Acc. (%)	H	R
Zieliński *et al*. [5]	CNN+SVM	97.0	No	No
Talo [6]	ResNet-50	99.2	No	No
Nasip and Zengin [8]	DenseNet201+SVM	99.78	No	No
Visitsattaponge *et al*. [9]	Big Transfer	99.11	No	No
Mai and Ishibashi [10]	DS-CNN	96.28	No	No
Chaudhari *et al*. [11]	CapsuleNet	95.08	No	No
AFN [7]	Ensemble+CMFF	96.8	No	Yes
HieroBact	AFN+HC-Loss	**97.2**	Yes	Yes

H: hierarchical outputs; R: robustness under image degradation reported. Differences in data splitting and augmentation mean that accuracy values are contextual rather than strictly head-to-head.

3.5. Robustness under image degradation

Under combined Gaussian noise and blur, HieroBact retained 89.3% species accuracy and 96.5% Gram accuracy, compared with 88.1% species accuracy for the flat baseline. This controlled stress test supports robustness to simulated degradation but does not replace external multi-site validation.

3.6. Discussion

The species-level accuracy advantage of HieroBact increases under degradation (+1.2 pp combined vs +0.4 pp clean), suggesting that the hierarchical auxiliary tasks provide implicit regularisation that benefits generalisation under distributional shift. This experiment does not replace multi-site staining validation, but it does provide a controlled stress test for blur, colour variation, and noise that commonly appears when microscopy images are captured under variable preparation and acquisition conditions.

From a clinical standpoint, a model errs may matter more than how often it errs. Prescribing G+-directed therapy for a G- pathogen, or vice versa, is not a minor discrepancy; it is a pharmacological mismatch that can precipitate treatment failure, expose patients to avoidable adverse effects, and contribute to resistance selection pressure [4]. HieroBact addresses this head-on: 94.3% of its species' misidentifications remain within the correct Gram group, leaving the most consequential clinical decision unaffected. We stress, however, that these results come from a controlled single-laboratory benchmark with standardised staining; robust validation on multi-site clinical material under realistic preparation variability is an essential prerequisite for any diagnostic deployment.

HC-Loss induces an emergent coarse-to-fine learning trajectory by construction. Because $\gamma(t) = \lambda_3 \cdot \hat{p}_g(y_g) \cdot \hat{p}_m(y_m)$, any collapse in Gram confidence (for example, after a particularly challenging augmented batch) automatically reduces $\gamma(t)$, dampening species gradients until the coarser representations recover their reliability. This self-regulation is absent from both fixed-weight MTL [28] and uncertainty-based weighting [29], which cannot account for the semantic ordering of tasks.

The framework is also compatible with practical diagnostic workflows, but not as a stand-alone clinical decision maker. A realistic deployment would use HieroBact after smear imaging and before confirmatory culture or molecular testing, presenting Gram reaction, morphology, species probability, and coherence flags to a microbiologist for review. The current three-backbone configuration is appropriate for offline or GPU-assisted laboratory inference; low-resource or telemedicine settings would require compression by backbone replacement, pruning, quantisation, or knowledge distillation. In this respect, HieroBact should be viewed as a taxonomy-aware modelling framework whose clinical value depends on future validation, not as a completed diagnostic device.

The study is limited by the small single-protocol dataset, imbalanced Gram and morphology groups, singleton vibrio and filamentous/spirilla categories, missing complete out-of-fold morphology and species vectors, and the 47.2M-parameter ensemble. External multicentre validation, calibrated uncertainty, explainability, and hardware-specific throughput measurements are required before clinical workflow integration.

4. CONCLUSION

This paper introduced HieroBact, a hierarchical MTL framework that predicts Gram reaction, cellular morphology, and species identity from Gram-stained bacterial images. The findings show that hierarchy-aware training preserves strong recognition performance while substantially reducing Gram-

incoherent species errors. Its principal contribution is therefore a more clinically coherent and interpretable error structure rather than a marginal increase in species accuracy alone.

The study is limited by the small, single-benchmark dataset, standardised staining conditions, imbalanced hierarchical groups, and singleton morphology classes. Future work should validate the framework prospectively on larger multicentre clinical datasets, test diverse staining protocols and imaging devices, and add clinician-facing explanations that connect image regions with Gram, morphology, and species decisions. Model compression, hardware latency measurement, and integration into real-time diagnostic microscopy workflows should follow only after those external validation and explainability steps.

FUNDING INFORMATION

Politeknik Negeri Banyuwangi provided an institutional research grant for this study. The authors gratefully acknowledge its financial and institutional support.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Sepyan Purnama Kristanto	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	
Moh. Erdda Habiby	✓	✓	✓		✓			✓	✓	✓				
Lutfi Hakim			✓	✓		✓	✓	✓	✓	✓	✓			
Dianni Yusuf				✓	✓	✓	✓	✓		✓	✓	✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest. The authors declare no known competing financial interests or personal relationships that could have appeared to influence this work.

DATA AVAILABILITY

DIBaS-Tax annotations, data splits, training scripts, and model-configuration details are available from the corresponding author (S.P.K.) upon reasonable request. The original DIBaS images are publicly available at http://misztal.edu.pl/software/databases/dibas/.

REFERENCES

[1] World Health Organization, "Diarrhoeal disease," *WHO Fact Sheet*, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diarrhoeal-disease.
[2] T. J. Beveridge, "Use of the Gram stain in microbiology," *Biotechnic & Histochemistry*, vol. 76, no. 3, pp. 111–118, 2001, [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/11475313/.
[3] P. R. Murray, K. S. Rosenthal, and M. A. Pfaller, *Medical microbiology*, 9th ed. Amsterdam: Elsevier, 2020.
[4] C. L. Ventola, "The antibiotic resistance crisis: part 1: causes and threats," *Pharmacy and Therapeutics*, vol. 40, no. 4, pp. 277–283, 2015.
[5] 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, 2017, doi: 10.1371/journal.pone.0184554.
[6] M. Talo, "An Automated Deep Learning Approach for Bacterial Image Classification," 2019, *arXiv*. doi: 10.48550/ARXIV.1912.08765.
[7] S. P. Kristanto, L. Hakim, and Hermansyah, "AquaFusionNet: Lightweight VisionSensor Fusion Framework for Real-Time Pathogen Detection and Water Quality Anomaly Prediction on Edge Devices," 2025, *arXiv*. doi: 10.48550/ARXIV.2512.06848.
[8] Ö. F. Nasip and K. Zengin, "A hybrid model: multiple feature selection approach using transfer learning for bacteria classification," *Traitement du Signal*, vol. 39, no. 6, pp. 2123–2131, Dec. 2022, doi: 10.18280/ts.390625.
[9] S. Visitsattaponge, 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.

[10] D.-T. Mai and K. Ishibashi, "Small-scale depthwise separable convolutional neural networks for bacteria classification," *Electronics*, vol. 10, no. 23, p. 3005, Dec. 2021, doi: 10.3390/electronics10233005.
[11] J. P. Chaudhari, H. Mewada, A. V. Patel, and K. Mahant, "Automated bacteria genera classification using histogram-oriented optimized capsule network," *Engineering Science and Technology, an International Journal*, vol. 46, 2023, doi: 10.1016/j.jestch.2023.101500.
[12] S. V, K. Prasad, C. Mukhopadhyay, and B. Banerjee, "Vision transformer based bacteria classification model for Gram-stained direct smear images," *Multimedia Tools and Applications*, vol. 84, no. 19, pp. 20289–20309, 2025, doi: 10.1007/s11042-025-20742-0.
[13] Z. Ge, D. Mahapatra, X. Chang, Z. Chen, L. Chi, and H. Lu, "Improving multi-label chest X-ray disease diagnosis by exploiting disease and health labels dependencies," *Multimedia Tools and Applications*, vol. 79, no. 21–22, pp. 14889–14902, 2020, doi: 10.1007/s11042-019-08260-2.
[14] 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 November, 2023, doi: 10.1371/journal.pcbi.1011181.
[15] M. Zachariou, O. Arandjelović, W. Sabiiti, B. Mtafya, and D. Sloan, "Tuberculosis bacteria detection and counting in fluorescence microscopy images using a multi-stage deep learning pipeline," *Information (Switzerland)*, vol. 13, no. 2, 2022, doi: 10.3390/info13020096.
[16] D. M. Wonohadidjojo and L. Handayani, "Classification of Gram-stained microscopy images of bacteria cultured on multiple media using frozen and fine-tuned ResNet-50 transfer learning: a comparative study," *Teknika*, vol. 15, no. 2, pp. 357–369, Jul. 2026, doi: 10.34148/teknika.v15i2.1507.
[17] K. B. Sanjaya and D. M. Wonohadidjojo, "Detection and classification of Gram-stained bacteria in microscopic images using YOLOV8 with CBAM," *JIKO (Jurnal Informatika dan Komputer)*, vol. 8, no. 3, pp. 275–284, Dec. 2025, doi: 10.33387/jiko.v8i3.10891.
[18] D. Li, Y. Zhu, A. Mehmood, Y. Liu, X. Qin, and Q. Dong, "Intelligent identification of foodborne pathogenic bacteria by self-transfer deep learning and ensemble prediction based on single-cell Raman spectrum," *Talanta*, vol. 285, p. 127268, Apr. 2025, doi: 10.1016/j.talanta.2024.127268.
[19] Q. u. Ain and S. Asif, "A novel ensemble approach with deep transfer learning for accurate identification of foodborne bacteria from hyperspectral microscopy," *Computational Biology and Chemistry*, vol. 113, p. 108238, Dec. 2024, doi: 10.1016/j.compbiolchem.2024.108238.
[20] Z. Sun, Z. Wang, and M. Jiang, "RamanCluster: a deep clustering-based framework for unsupervised Raman spectral identification of pathogenic bacteria," *Talanta*, vol. 275, p. 126076, Aug. 2024, doi: 10.1016/j.talanta.2024.126076.
[21] Y. Zou, Y. Li, F. Zhang, Y. Ge, W. Wang, and M. Chen, "Rapid identification of Litopenaeus vannamei pathogenic bacteria: a combined approach using surface-enhanced Raman spectroscopy (SERS) and deep learning," *Analytical and Bioanalytical Chemistry*, vol. 417, no. 20, pp. 4587–4603, Aug. 2025, doi: 10.1007/s00216-025-05974-1.
[22] C.-S. Ho *et al.*, "Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning," *Nature Communications*, vol. 10, no. 1, p. 4927, Oct. 2019, doi: 10.1038/s41467-019-12898-9.
[23] H. Chen, S. Miao, D. Xu, G. D. Hager, and A. P. Harrison, "Deep hiearchical multi-label classification applied to chest X-ray abnormality taxonomies," *Medical Image Analysis*, vol. 66, p. 101811, Dec. 2020, doi: 10.1016/j.media.2020.101811.
[24] K. T. Noor, A. Robles-Kelly, L. Y. Zhang, M. R. Bouadjenek, and W. Luo, "A consistency-aware deep capsule network for hierarchical multi-label image classification," *Neurocomputing*, vol. 604, p. 128376, Nov. 2024, doi: 10.1016/j.neucom.2024.128376.
[25] D. Muhammad and M. Bendechache, "Unveiling the black box: a systematic review of Explainable Artificial Intelligence in medical image analysis," *Computational and Structural Biotechnology Journal*, vol. 24, pp. 542–560, Dec. 2024, doi: 10.1016/j.csbj.2024.08.005.
[26] B. H. M. v. d. Velden, H. J. Kuijf, K. G. A. Gilhuijs, and M. A. Viergever, "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis," *Medical Image Analysis*, vol. 79, p. 102470, Jul. 2022, doi: 10.1016/j.media.2022.102470.
[27] D. J. Brenner, N. R. Krieg, J. T. Staley, and G. M. Garrity, Eds., *Bergey's manual® of systematic bacteriology*. Boston, MA: Springer US, 2005.
[28] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, no. 1, pp. 41–75, Jul. 1997, doi: 10.1023/A:1007379606734.
[29] R. Cipolla, Y. Gal, and A. Kendall, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 7482–7491, doi: 10.1109/CVPR.2018.00781.
[30] Z. Chen, V. Badrinarayanan, C.-Y. Lee, and A. Rabinovich, "GradNorm: Gradient Normalization for Adaptive Loss Balancing in Deep Multitask Networks," *Proceedings of the International Conference on Machine Learning*, pp. 794–803, 2018.
[31] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *ACM International Conference Proceeding Series*, Jun. 2009, vol. 382, pp. 41–48, doi: 10.1145/1553374.1553380.
[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, vol. 2016-Decem, pp. 770–778, doi: 10.1109/CVPR.2016.90.
[33] M. Tan and Q. Le, "EfficientNet: rethinking model scaling for convolutional neural networks," *Proceedings of Machine Learning Research*, vol. 97, pp. 6105–6114, 2019.
[34] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE*, pp. 2261–2269, Jul. 2017. doi: 10.1109/cvpr.2017.243.
[35] T. Ridnik, E. Ben-Baruch, A. Noy, and L. Zelnik-Manor, "ImageNet-21K Pretraining for the Masses," 2021, *arXiv*. doi: 10.48550/ARXIV.2104.10972.
[36] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," 2017, *arXiv*. doi: 10.48550/ARXIV.1711.05101.
[37] I. Loshchilov and F. Hutter, "SGDR: Stochastic Gradient Descent with Warm Restarts," 2016, *arXiv*. doi: 10.48550/ARXIV.1608.03983.
[38] Q. McNemar, "Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, Jun. 1947, doi: 10.1007/bf02295996.
[39] K. D. Apostolidis and G. A. Papakostas, "A survey on adversarial deep learning robustness in medical image analysis," *Electronics*, vol. 10, no. 17, p. 2132, Sep. 2021, doi: 10.3390/electronics10172132.
[40] R. Guo, Y. Xu, A. Tompkins, M. Pagnucco, and Y. Song, "Multi-degradation-adaptation network for fundus image enhancement with degradation representation learning," *Medical Image Analysis*, vol. 97, p. 103273, Oct. 2024, doi: 10.1016/j.media.2024.103273.
[41] R. Setayeshi, J. Vahidi, E. Kozegar, and T. Tan, "An end-to-end multi-task deep learning framework for bronchoscopy image

classification," *Multimedia Systems*, vol. 30, no. 6, p. 361, Dec. 2024, doi: 10.1007/s00530-024-01579-3.
[42] S. Yu, H. Li, X. Li, Y. V. Fu, and F. Liu, "Classification of pathogens by Raman spectroscopy combined with generative adversarial networks," *Science of The Total Environment*, vol. 726, p. 138477, Jul. 2020, doi: 10.1016/j.scitotenv.2020.138477.
[43] B. Liu *et al.*, "Laser tweezers Raman spectroscopy combined with deep learning to classify marine bacteria," *Talanta*, vol. 244, p. 123383, Jul. 2022, doi: 10.1016/j.talanta.2022.123383.

BIOGRAPHIES OF AUTHORS



Sepyan Purnama Kristanto received Bachelor's degrees from Edith Cowan University and Universitas Bakti Indonesia, and the M.Kom. degree in Information Technology from Institut Sains dan Teknologi Terapan Surabaya. He is a lecturer at Politeknik Negeri Banyuwangi. His research interests include machine learning, computer vision, human-computer interaction, and generative AI. He can be contacted at email: sepyan@poliwangi.ac.id.



Moh. Erdda Habiby is a faculty member at the School of Computer Science, BINUS University. His research interests include web development, data analytics, and artificial intelligence. He can be contacted at email: moh.erdda@binus.ac.id.



Lutfi Hakim is a lecturer at Politeknik Negeri Banyuwangi. His research interests include intelligent systems, pattern recognition, and application development. He can be contacted at email: lutfi@poliwangi.ac.id.



Dianni Yusuf is a lecturer at Politeknik Negeri Banyuwangi. Her research interests include software engineering, software development, and applied artificial intelligence. She can be contacted at email: dianni@poliwangi.ac.id.