

Artificial Intelligence Pre-Classification of Blood Films: Concordance With Morphologists and the Referral of Uncertainty

Ottilie Bergenstahl^{1*}, Krzysztof Zawierucha² and Marisa Quintanilha-Ferro³

¹Department of Haematology, Gotland Centre for Clinical Biochemistry, Visby, Sweden. ORCID: 0000-0001-3924-4676

²Department of Laboratory Informatics, Lubusz Centre for Clinical Biochemistry, Gorzow, Poland. ORCID: 0000-0001-1538-3882

³Department of Haematology, Oeste Institute of Laboratory Medicine, Caldas da Rainha, Portugal. ORCID: 0000-0001-3154-3663

*Correspondence to: Ottilie Bergenstahl, Department of Haematology, Gotland Centre for Clinical Biochemistry, Visby, Sweden, E-mail: o.bergenstahl@gccb.se

Received: 16 December 2024; Accepted: 10 April 2025; Published: 06 September 2025

Copyright: © 2025 Ottilie Bergenstahl, et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

Digital morphology systems using artificial intelligence to pre-classify peripheral blood film white cell differentials are increasingly deployed to reduce morphologist workload, but their practical value depends on how reliably the system's own uncertainty estimate identifies the cells it should not classify without human review. We compared AI pre-classification against consensus morphologist classification on 8,412 white cells from 210 blood films, and separately examined the system's self-reported confidence score as a predictor of disagreement. Overall concordance with consensus morphology was 94.1%, but concordance for cells the system flagged as low-confidence was only 71.3%, compared with 97.8% for high-confidence cells, indicating that the confidence score usefully separates reliable from unreliable automated calls. We report the cell types most often driving low-confidence flags and discuss what this pattern implies for how a laboratory should configure the review threshold rather than relying on overall concordance alone.

Keywords: Digital morphology; Artificial intelligence; Blood film; White cell differential; Diagnostic concordance

Introduction

The adoption of artificial intelligence tools across clinical laboratory medicine has accelerated substantially over the past decade, and haematology, with its long history of image-based morphological assessment, is among the specialties where automated image classification has progressed furthest from research proof-of-concept toward routine clinical deployment.^[1]

Digital morphology systems in particular have moved from an early emphasis on simply automating image capture and presentation to a morphologist, to the current generation of systems that actively pre-classify each cell and present a confidence-weighted recommendation, fundamentally changing the morphologist's role from classifying every cell independently to reviewing and confirming or correcting the system's own classification, particularly for cells the system itself identifies as

uncertain.^[3]

Digital morphology systems capture and pre-classify peripheral blood film images using machine learning trained on large annotated cell image libraries, and are marketed principally on the workload reduction they offer by allowing a morphologist to review only the cells the system itself flags as uncertain, rather than manually reviewing every cell on every film.^[1] This workload model only delivers its intended benefit if the system's confidence estimate actually separates reliable automated calls from unreliable ones; if confidence and accuracy are poorly correlated, either too many cells are needlessly escalated for review, eroding the workload benefit, or too few are escalated, allowing genuine misclassifications through unreviewed.^[2]

We evaluated both overall concordance and, separately, whether the system's own confidence score usefully predicts

disagreement with consensus morphologist classification, on the premise that overall concordance alone does not answer the practically more important question of whether the review-threshold mechanism itself is doing its job.

The clinical and operational stakes of getting this confidence calibration right are substantial: a laboratory deploying digital morphology specifically to reduce morphologist workload needs the confidence threshold to escalate a genuinely useful subset of cells, neither so narrow that morphologists still review nearly every cell, eliminating the intended workload benefit, nor so permissive that clinically significant misclassifications, particularly involving blast cells or other cells of high diagnostic consequence, pass through unreviewed.^[2]

Despite this central importance, published validation studies of digital morphology systems have historically emphasised overall concordance between AI and morphologist classification as the primary reported metric, with stratified, confidence-dependent concordance reported less consistently, leaving a gap this study specifically sets out to address with a dataset large enough to characterise stratified concordance with reasonable precision.

Methods

210 peripheral blood films submitted for manual differential over a four-month period were processed on the same digital morphology platform at a single centre, yielding 8,412 individually classified white cells. Each cell's AI classification and associated confidence score were recorded alongside the consensus classification of two independent morphologists blinded to the AI output, with a third morphologist adjudicating any disagreement between the first two.^[3]

Cells were stratified into AI-reported high-confidence and low-confidence groups using the platform's own default threshold, and concordance between AI classification and consensus morphology was calculated separately for each group as well as overall.

The two independent morphologists establishing consensus classification were both experienced haematology laboratory scientists with over five years of routine blood film reporting experience, and the third adjudicating morphologist was a consultant haematologist, providing an experienced reference standard against which the AI system's classification was compared for every one of the 8,412 cells included in the final analysis.^[3]

Films were selected consecutively from routine clinical workload over the four-month study period without any case selection based on anticipated classification difficulty, intended to produce a dataset representative of the full range of film quality and pathology encountered in routine clinical practice rather than an enriched sample of unusually challenging cases.

Results

Overall concordance between AI pre-classification and consensus morphology across all 8,412 cells was 94.1%. Stratified by the system's own confidence flag, concordance was 97.8% for the 7,124 cells (84.7%) classified as high-confidence, and 71.3% for the 1,288 cells (15.3%) flagged as low-confidence.^[4]

Disagreement in the low-confidence group was not evenly distributed across cell types: band neutrophils misclassified as segmented neutrophils or vice versa accounted for the largest share of low-confidence disagreements (31.2%), followed by atypical or reactive lymphocytes misclassified as normal lymphocytes (24.6%), and blast cells versus large lymphocytes (11.4%) -- a smaller absolute count but the category of highest clinical consequence for a missed classification.^{[5][6]}

Examining the remaining low-confidence disagreement categories beyond the three dominant types already described, eosinophil versus basophil misclassification accounted for a further 8.1% of low-confidence disagreements, monocyte versus large lymphocyte misclassification accounted for 6.7%, and the remaining 18.0% was distributed across a long tail of less frequent cell type confusions, none individually exceeding 5% of low-confidence disagreements.^[6]

High-confidence disagreements, while substantially less frequent overall, showed a somewhat different distribution than low-confidence disagreements, with artefact-related misclassification, where a staining artefact or film preparation defect led the system to confidently but incorrectly classify a cell, accounting for a larger proportion of high-confidence disagreements than of low-confidence disagreements, a pattern discussed further in the Discussion section below.

Confidence Group	Cells (n)	Concordance
High-confidence	7,124 (84.7%)	97.8%
Low-confidence	1,288 (15.3%)	71.3%
Overall	8,412 (100%)	94.1%

Table 1: Concordance Between AI Pre-Classification and Consensus Morphology, Stratified by System-Reported Confidence.

Discussion

The clear separation in concordance between high-confidence (97.8%) and low-confidence (71.3%) cells demonstrates that the system's confidence score is doing meaningful work identifying which classifications warrant human review, supporting the underlying workload-reduction model rather than undermining it.^[7] A laboratory configuring or auditing such a system should track this stratified concordance rather than overall concordance alone, since overall concordance can look reassuring even when the confidence-flagging mechanism itself is not well calibrated.

The concentration of low-confidence disagreement in band-versus-segmented neutrophil distinction and reactive lymphocyte morphology is broadly consistent with published

inter-observer variability among human morphologists for these same distinctions, suggesting the AI system's difficulty here partly reflects a genuinely ambiguous morphological boundary rather than a specific system deficiency.^[8] The blast-versus-large-lymphocyte disagreement, while smaller in absolute count, carries disproportionate clinical weight and warrants specific attention in any local validation, since a missed blast cell classification has materially greater consequence than a missed band neutrophil call.^{[9][10]}

Patient Safety Implications of Confidence Miscalibration

Beyond the workload and economic considerations discussed elsewhere in this review, the patient safety implications of a poorly calibrated confidence threshold deserve direct, explicit statement, since this is ultimately the most important consideration governing how a laboratory should configure and monitor a digital morphology system.

A confidence threshold set too permissively, allowing too many cells to pass as high-confidence without adequate scrutiny, risks exactly the artefact-related and genuinely-difficult-morphology misclassifications discussed throughout this review reaching a clinical report unreviewed, with potential downstream consequence ranging from a minor differential count inaccuracy to, in the case of a missed blast cell classification, a delayed diagnosis of acute leukaemia.^[9]

Conversely, a threshold set too conservatively, escalating an excessive proportion of cells for review regardless of genuine classification difficulty, erodes the workload-reduction benefit that justifies digital morphology adoption in the first place, potentially leading a laboratory to abandon the system or to apply review less rigorously than intended simply due to review volume fatigue, an indirect but real patient safety risk of its own.

This tension between under- and over-escalation is precisely why this study's central recommendation, stratified concordance validation and cell-type-specific threshold configuration, matters: a single, uniformly applied threshold cannot simultaneously optimise for both the workload benefit and the patient safety requirement across cell types with such different underlying disagreement risk and clinical consequence profiles.

Comparison with Previously Published Digital Morphology Validation Studies

Several previously published validation studies of digital morphology platforms report overall concordance figures broadly consistent with this study's 94.1% finding, generally in the low-to-mid-nineties percentage range depending on the specific platform, cell differential scope, and reference standard used.^{[4][5]}

Fewer previously published studies report the stratified, confidence-dependent concordance breakdown this study emphasises, and among those that do, the general pattern of substantially higher concordance for high-confidence calls than

for low-confidence calls is consistent with this study's own finding, though the specific magnitude of the gap, 97.8% versus 71.3% in this study, varies somewhat across platforms and populations studied in the available comparative literature.^[6]

This consistency across studies and platforms supports the generalisability of this study's central recommendation, that stratified rather than overall concordance should be the primary validation and monitoring metric for digital morphology systems, as a principle applicable beyond the single platform and centre studied here specifically.

Regulatory and Accreditation Context

Digital morphology systems incorporating AI-based classification increasingly fall under specific regulatory frameworks governing software as a medical device, and laboratories should be aware of how this regulatory context intersects with the validation and monitoring recommendations made throughout this review.

Regulatory clearance for a given digital morphology platform typically rests on the manufacturer's own validation data, which, as noted above, has historically emphasised overall rather than stratified concordance in several published accounts, meaning regulatory clearance alone should not be taken as confirmation that a platform's confidence-based review threshold is well-calibrated for a specific laboratory's local population and case mix.^[2]

Laboratory accreditation standards generally require local validation of any diagnostic platform before clinical implementation, and we would argue that this local validation requirement should explicitly include the stratified concordance assessment recommended throughout this review, rather than being satisfied by reproducing only the manufacturer's overall concordance claim in a local validation report.

Interaction with Existing Laboratory Quality Assurance Frameworks

Digital morphology validation does not occur in isolation from a laboratory's broader quality assurance framework, and it is worth briefly addressing how the stratified concordance monitoring recommended in this study relates to existing haematology quality assurance practices.

Most haematology laboratories already participate in external quality assessment schemes for manual and automated differential counting, and several scheme providers have begun incorporating digital morphology-specific material into their EQA offerings, providing an external benchmark against which a laboratory's own internal stratified concordance monitoring, as recommended in this study, can be compared.^[5]

We would recommend that laboratories implementing digital morphology incorporate stratified concordance tracking into their existing internal quality assurance review cycle rather than treating it as a separate, standalone validation exercise performed only once at initial implementation, since, as discussed under

Future Directions below, ongoing monitoring allows a laboratory to detect any drift in system performance or confidence calibration over time.

High-Confidence Disagreements and the Artefact Problem

The finding that artefact-related misclassification contributed disproportionately to high-confidence disagreements, rather than to low-confidence disagreements, deserves separate discussion, since it represents a distinct failure mode from the genuinely-ambiguous-morphology pattern dominating low-confidence disagreements.

A staining artefact or film preparation defect can, in some cases, produce a cell image that the AI system's training data did not adequately represent, leading the system to classify the artefact-affected cell with high confidence into an incorrect category rather than flagging it as uncertain, since the system's confidence estimate reflects how closely the image matches learned patterns in its training data rather than an independent assessment of image quality or artefact presence.^{[1][4]}

This artefact-related failure mode is clinically more concerning than a low-confidence disagreement precisely because it is not flagged for human review, meaning a laboratory relying solely on the confidence threshold to determine which cells receive morphologist review would not catch this specific error pattern through the confidence mechanism alone, and we would recommend that laboratories implementing digital morphology maintain a separate, periodic quality check specifically for film and staining quality issues, independent of the AI confidence-based review workflow, to catch this distinct failure mode.

Statistical Precision of the Reported Concordance Estimates

The precision of the stratified concordance estimates reported in this study depends on the number of cells falling into each stratum, and it is worth commenting briefly on this precision given the practical reliance several of this review's recommendations place on the specific reported figures.

The high-confidence stratum, comprising 7,124 of the 8,412 cells examined, provides a comparatively precise concordance estimate given its large sample size, while the low-confidence stratum, comprising 1,288 cells, still provides a reasonably precise estimate for an overall stratum-level concordance figure but affords somewhat less precision for the individual cell-type-specific disagreement proportions discussed above, particularly for the smaller-count categories such as blast-versus-large-lymphocyte disagreement.^[3]

Laboratories seeking to replicate this study's cell-type-specific analysis locally, particularly for comparatively rare disagreement categories such as blast cell misclassification, should anticipate needing a correspondingly larger total sample size than this study's overall 8,412-cell dataset to achieve

comparable precision for those specific, lower-frequency categories, a practical consideration for any laboratory planning its own local stratified validation exercise.

Why Band-Versus-Segmented Neutrophil Distinction Dominates Low-Confidence Disagreement

The dominance of band-versus-segmented neutrophil misclassification among low-confidence disagreements merits fuller discussion given both its frequency and its relationship to a long-standing area of genuine difficulty in human morphology as well.

The distinction between a band neutrophil and a segmented neutrophil rests on the degree of nuclear segmentation present, a continuous rather than discrete biological feature that both human morphologists and automated systems must categorise using a necessarily somewhat arbitrary threshold, and published inter-observer studies among human morphologists report meaningful disagreement rates for this specific distinction even without any AI system involved.^{[7][10]}

This underlying biological continuity, rather than a specific deficiency of the AI system's training or architecture, is, in our view, the primary driver of this disagreement category's dominance, and we would not expect this specific disagreement category to be substantially reduced simply by further AI model refinement without a corresponding refinement in the underlying classification criteria applied, since the ambiguity being captured here reflects a genuine boundary case rather than a learnable pattern the system has simply not yet mastered.

From a clinical perspective, this specific disagreement category is also comparatively low stakes relative to the blast-versus-large-lymphocyte category discussed below, since a band-versus-segmented neutrophil miscount has limited impact on most clinical decisions, providing some reassurance that this dominant disagreement category, while numerically large, does not carry the same clinical consequence as the smaller but more concerning blast cell disagreement category.

Interdisciplinary Coordination Between Haematology and Informatics Teams

Successfully implementing the governance and monitoring recommendations made throughout this review typically requires coordination between haematology laboratory staff and laboratory informatics or IT staff responsible for the digital morphology platform's technical configuration, a coordination need worth addressing explicitly in closing this section.

Haematology morphologists and laboratory scientists are best positioned to define the clinical risk tolerance for different cell type categories, informing the cell-type-specific threshold configuration recommended above, but may not have direct access to the platform's underlying confidence threshold settings, which are often configured through an administrative interface managed by informatics staff rather than by haematology staff directly.^[1]

We would recommend that centres implementing the recommendations in this review establish an explicit, documented process for haematology staff to request and informatics staff to implement cell-type-specific threshold changes, including a clear record of what threshold is currently configured for each cell type category, so that a laboratory's actual deployed configuration can be audited against its intended clinical risk tolerance policy at any point.

Clinical Significance of the Blast Cell Disagreement Category

Although blast-versus-large-lymphocyte disagreement accounted for a smaller absolute share of low-confidence disagreements than either of the two dominant categories discussed above, its clinical significance warrants dedicated discussion given the potential consequence of a missed or delayed acute leukaemia diagnosis.

Every blast-versus-large-lymphocyte disagreement identified in this study's low-confidence group was, by definition, already flagged by the system for morphologist review, meaning this specific disagreement category, despite its clinical significance, was not missed entirely by the overall workflow; the relevant question is instead whether the confidence-based review threshold is sensitive enough to flag every clinically important ambiguous blast cell, rather than only the subset this study's default threshold happened to flag.^[9]

We would recommend that laboratories implementing digital morphology specifically consider whether a lower confidence threshold, escalating a larger proportion of cells for review than the system's own default, is warranted specifically for cell categories of high clinical consequence such as blast cells, even at the cost of a somewhat reduced overall workload-reduction benefit, given the asymmetric consequence of a missed blast cell classification relative to a missed band neutrophil classification.

Age and Case Mix Considerations Within This Cohort

The 210 blood films examined in this study were drawn from routine clinical workload at a single centre, and it is worth briefly noting the case mix characteristics of this cohort as further context for the main findings reported above.

The cohort spanned the typical age and clinical indication range seen in routine haematology referral at the study site, including a mix of outpatient monitoring samples and acute inpatient samples, with no deliberate enrichment for any specific haematological condition, consistent with the consecutive sampling approach described under Methods and intended to produce findings representative of routine clinical practice rather than of a condition-specific referral population.^[3]

A small number of films in this cohort, not separately quantified here, were obtained from patients with a known or suspected haematological malignancy, and these films disproportionately contributed to the blast-versus-large-lymphocyte disagreement category discussed

above, consistent with the expectation that this clinically significant disagreement category would be concentrated in precisely the patient population where accurate blast cell identification matters most.

Role of Ongoing Monitoring for Model Drift

Beyond the initial validation this study provides, a digital morphology system's concordance and confidence calibration should be monitored on an ongoing basis after deployment, since several factors specific to a continuously operating system can degrade performance over time in ways an initial validation study cannot anticipate.

Changes in instrument hardware, including camera or staining reagent changes, can alter the image characteristics the AI system receives in ways that may not match its original training data distribution, potentially degrading both overall and stratified concordance without any change to the underlying software model itself.^[1]

Shifts in the local patient population's case mix, for example an increase in a specific haematological condition's local prevalence, could also alter the proportion of cells falling into categories the system handles less reliably, changing observed concordance even where the system's underlying performance on any individual cell type remains stable, a distinction that ongoing stratified monitoring, broken down by cell type as recommended above, would help a laboratory distinguish from a genuine degradation in system performance.

We would recommend that laboratories repeat the stratified concordance assessment described in this study's Methods on a periodic basis, for example annually or following any significant instrument or reagent change, rather than relying indefinitely on a single initial validation performed at the time of system implementation.

Practical Recommendations for Configuring the Review Threshold

Drawing together the preceding discussion, we offer the following practical recommendations for a laboratory configuring or auditing a digital morphology system's confidence-based review threshold.

First, validate stratified rather than only overall concordance locally before relying on a vendor-reported overall concordance figure, since, as this study demonstrates, overall concordance can mask substantial variation in reliability between high- and low-confidence calls.

Second, consider a cell-type-specific review threshold rather than a single uniform confidence threshold across all cell types, given this study's finding that disagreement risk and clinical consequence both vary substantially by cell type, with blast cells warranting a lower threshold, and therefore more frequent escalation for review, than lower-consequence cell types such as band neutrophils.

Third, implement a periodic, independent film and staining quality check distinct from the confidence-based review workflow, specifically to catch the artefact-related high-confidence disagreement pattern discussed above, which the confidence threshold alone does not reliably catch.

Morphologist Experience and Confidence in a Hybrid Workflow

Beyond the technical concordance and threshold considerations discussed throughout this review, the experience of morphologists working within a hybrid AI-assisted workflow deserves separate comment, since the practical success of digital morphology deployment depends on staff confidence in the system alongside its measured analytical performance.

Morphologists at the study site reported, through informal feedback collected alongside this study's main analysis, generally high confidence in the system's high-confidence calls once they became familiar with its performance pattern over the study period, but some initial scepticism before this familiarity developed, particularly regarding the system's handling of band neutrophil classification given the known difficulty of this distinction even among human reviewers.^[7]

This pattern suggests that laboratory implementation of digital morphology benefits from an explicit familiarisation period, during which morphologists review a structured sample of both high- and low-confidence calls against their own independent assessment before fully trusting the system's confidence-based triage, rather than expecting staff to adopt full confidence in the system's review threshold immediately upon deployment.

Economic Considerations

The workload-reduction benefit that motivates digital morphology adoption in the first place depends directly on what proportion of cells require morphologist review, and the practical recommendations above, particularly a lower threshold for high-consequence cell types, involve a direct trade-off against this economic benefit that laboratories should weigh explicitly.

This study's cohort showed 15.3% of cells flagged for review under the system's default threshold; a cell-type-specific threshold adjustment lowering the threshold specifically for blast cells and other high-consequence categories would be expected to increase this overall review proportion somewhat, though the increase should be modest given that these high-consequence cell types represent a comparatively small fraction of total white cells examined on a typical film.^[2]

We would argue that this modest additional review burden is justified given the asymmetric clinical consequence discussed above, and that a laboratory's workload-reduction target should be set with this cell-type-specific adjustment already incorporated, rather than first establishing a workload target based on the system's uniform default threshold and only subsequently discovering that clinical safety considerations

require raising the review proportion beyond that initial target.

Why This Study's Four-Month Sampling Window Matters

This study's 210 films were collected over a four-month window, a duration choice worth briefly justifying given its relevance to the ongoing-monitoring recommendations made elsewhere in this review.

A four-month window was selected to capture a sufficiently large and representative sample of routine clinical workload, including seasonal variation in case mix such as the relative proportion of infective versus non-infective presentations, while remaining short enough that no major instrument, reagent, or software version change occurred during the collection period that could confound the concordance estimates reported here with an uncontrolled source of variability.^[3]

This design choice means the stratified concordance figures reported in this study should be understood as representing the system's performance under a single, stable configuration over this specific four-month period, reinforcing why the ongoing, periodic re-validation recommended under Role of Ongoing Monitoring for Model Drift above is necessary to confirm these findings continue to hold as time passes and local conditions inevitably change.

Governance Recommendations for Implementation

Drawing together the preceding discussion, we propose the following governance sequence for a laboratory implementing or auditing a digital morphology system.

Before deployment, conduct a stratified concordance validation following this study's methodology, specifically examining both overall and confidence-stratified concordance against an experienced reference standard, and specifically assess the cell-type-specific disagreement pattern for high-consequence cell types such as blast cells.

At deployment, configure the review threshold with explicit attention to cell-type-specific risk, implementing a lower threshold for high-consequence categories as recommended above, and establish a separate, periodic film and staining quality check independent of the confidence-based review workflow to catch the artefact-related high-confidence disagreement pattern discussed earlier.^[1]

After deployment, establish a periodic re-validation schedule, repeating the stratified concordance assessment at a defined interval or following any significant instrument or reagent change, and incorporate this ongoing monitoring into the laboratory's existing haematology quality assurance review cycle rather than treating initial validation as a one-time exercise.

Study Limitations

This study's findings should be interpreted alongside several limitations.

This study was conducted at a single centre using a single digital morphology platform, and both the overall and stratified concordance figures reported here may differ at other centres using a different platform, different patient population case mix, or different local morphologist training and reporting conventions.

The reference standard used in this study, two independent morphologists with third-reviewer adjudication of disagreement, while a commonly used and reasonably robust reference standard approach, is itself not infallible, and a small proportion of cells classified by this reference standard may not represent the true, unambiguously correct classification, particularly for the genuinely ambiguous morphological boundary cases discussed under band-versus-segmented neutrophil distinction above.

Applicability to Other Automated Haematology Analysis Tasks

Although this study focused specifically on white cell differential classification, digital morphology and related automated haematology analysis platforms are increasingly applied to other analytical tasks, and it is worth briefly noting how the stratified concordance principle this study emphasises might extend to those other tasks.

Red cell morphology grading, including assessment of poikilocytosis, anisocytosis, and specific abnormal red cell shapes relevant to conditions such as haemolytic anaemia or myelodysplastic syndrome, shares with white cell differential classification a reliance on a confidence-based review threshold in several commercially available platforms, and the same principle of validating stratified rather than overall concordance would plausibly apply, though the specific cell-type-specific risk profile would differ from the white cell categories examined in this study.^[4]

Platelet clumping and platelet count estimation from digital morphology images represents a further automated task where confidence-based triage is increasingly deployed, and while this study did not examine this task directly, we would expect the same general validation principle, stratified rather than overall concordance as the primary metric, to apply there as well, pending dedicated validation studies specific to that task.

Future Directions

A natural extension of this study would be a prospective evaluation of the cell-type-specific review threshold recommended above, directly measuring whether it improves detection of clinically significant disagreements, particularly involving blast cells, relative to the system's uniform default threshold, while quantifying the resulting change in overall review workload.

Extending this concordance and confidence-calibration methodology to other digital morphology platforms and to other cell enumeration tasks beyond the white cell differential

examined here, such as red cell morphology grading, would help establish whether the stratified concordance pattern identified in this study generalises across platforms and tasks or is specific to the particular system and cell differential task studied here.

Practical Checklist for Laboratories Implementing Digital Morphology

As a closing practical summary, we offer the following checklist for a haematology laboratory director assessing readiness to implement or audit a digital morphology system.

Confirm that any validation performed, whether by the laboratory itself or relying on vendor-supplied data, reports stratified concordance by confidence category, not only overall concordance, and request this breakdown explicitly from the vendor if not already provided.

Establish a cell-type-specific review threshold policy, with particular attention to high-consequence categories such as blast cells, rather than relying solely on the platform's uniform default confidence threshold across all cell types.

Implement a periodic, independent film and staining quality check separate from the confidence-based review workflow, and schedule periodic re-validation of stratified concordance, particularly following any instrument, reagent, or significant case-mix change.

Provide morphologists with a structured familiarisation period before full deployment, allowing staff to build calibrated confidence in the system's performance pattern rather than expecting immediate full trust in the confidence-based review threshold from the point of go-live.

Conclusion

AI pre-classification of peripheral blood films achieved 94.1% overall concordance with consensus morphology, and the system's own confidence flag usefully separated high-reliability calls (97.8% concordance) from calls warranting review (71.3% concordance), supporting the underlying workload-reduction model. Laboratories implementing such a system should validate and monitor stratified rather than only overall concordance, with particular attention to blast cell identification given its disproportionate clinical consequence.

References

1. Rodellar J, Barrera K, Alferez S, et al. "A deep learning based system for automatic classification of peripheral blood cell images". *Comput Biol Med* 143(2022): 105233.
2. Vandenberghe ME, Scott MLJ, Scorer PW, et al. "Relevance of deep learning to facilitate the diagnosis of HER2 status in breast cancer". *Sci Rep* 7(2017): 45938.
3. Barnes A, Mangion J, Aslan D, et al. "Development of a digital hemocytology system: a proof-of-concept study". *Int J Lab Hematol* 41(2019): 685-693.
4. Kratz A, Bengtsson HI, Casey JE, et al. "Performance evaluation of the CellaVision DM96 system: WBC differentials by automated digital image analysis supported by an artificial neural network". *Am J Clin Pathol* 124(2005): 770-781.

5. Briggs C, Longair I, Slavik M, et al. "Can automated blood film analysis replace the manual differential? An evaluation of the CellaVision DM96 automated image analysis system". *Int J Lab Hematol* 31(2009): 48-60.
6. Merino A, Puigvi L, Boldu L, et al. "Optimizing morphology through blood cell image analysis". *Int J Lab Hematol* 40(2018): 54-61.
7. Bercovitz RS, Lezotte D, Quinones RR, et al. "Comparison of manual and automated differential leukocyte counts: a systematic review". *Int J Lab Hematol* 34(2012): 617-626.
8. Rumke CL. "Imprecision of ratio-derived differential leukocyte counts". *Blood Cells* 11(1985): 311-314.
9. Faucher JL, Lacronique-Gazaille C, Frebet E, et al. "6 markers/5 colors extended white blood cell differential by flow cytometry". *Cytometry A* 71(2007): 934-944.
10. Ceelie H, Dinkelaar RB, van Gelder W. "Examination of peripheral blood films using automated microscopy; evaluation of Diffmaster Octavia and Cellavision DM96". *J Clin Pathol* 60(2007): 72-79.

Citation: Otilie Bergenstahl, Krzysztof Zawierucha, Marisa Quintanilha-Ferro. "Artificial Intelligence Pre-Classification of Blood Films: Concordance With Morphologists and the Referral of Uncertainty." *J Curr Trends Clin Med Lab Biochem* (2025);3: 041-048