Research Square

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

Separating method failure from reference-data limits in automated thin-section petrography: palaeovolcanic rocks of southern Uzbekistan

Zokir D. Yoldoshev
yuldoshevz1995@gmail.com
Institute of Geology and Geophysics named after H.M. Abdullaev
Usmonjon D. Mamarozikov
University of Geological Sciences
Ahmadali H. Hamrayev
Institute of Geology and Geophysics named after H.M. Abdullaev
Nuriddin E. Saitov
Institute of Geology and Geophysics named after H.M. Abdullaev
Utkir A. Normatov
Institute of Geology and Geophysics named after H.M. Abdullaev
Utkir O. Ikromov
Institute of Geology and Geophysics named after H.M. Abdullaev
Zilola A. Keldibekova
Institute of Geology and Geophysics named after H.M. Abdullaev
Ulugbek H. Tukhtayev
University of Geological Sciences

Method Article

Keywords:

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11807649/v1

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

Additional Declarations: No competing interests reported.

Separating method failure from reference-data limits in automated thin-section petrography: palaeovolcanic rocks of southern Uzbekistan

Zokir D. Yoldoshev [1,*], Usmonjon D. Mamarozikov [2,1], Ahmadali H. Hamrayev [1], Nuriddin E. Saitov [1], Utkir A. Normatov [1], Utkir O. Ikromov [1], Zilola A. Keldibekova [1], Ulugbek H. Tukhtayev [2]

[1] Institute of Geology and Geophysics named after H.M. Abdullaev, 64G Olimlar Street, Tashkent 100041, Uzbekistan

[2] University of Geological Sciences, 64A Olimlar Street, Tashkent 100041, Uzbekistan

Corresponding author

Zokir D. Yoldoshev, Institute of Geology and Geophysics named after H.M. Abdullaev, 64G Olimlar Street, Tashkent 100041, Uzbekistan. E-mail: yuldoshevz1995@gmail.com. ORCID: 0009-0004-2428-9002.

ORCID

Zokir D. Yoldoshev: 0009-0004-2428-9002. ORCID identifiers were not available for the remaining authors at the time of submission.

Author contributions

Zokir D. Yoldoshev: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. Usmonjon D. Mamarozikov: Conceptualization, Supervision, Validation, Writing – review & editing. Ahmadali H. Hamrayev: Investigation, Resources, Writing – review & editing. Nuriddin E. Saitov: Investigation, Resources, Validation. Utkir A. Normatov: Investigation, Data curation. Zilola A. Keldibekova: Data curation, Writing – review & editing. Ulugbek H. Tukhtayev and Utkir O. Ikromov: Investigation, Resources. All authors read and approved the final manuscript.

Acknowledgements

The authors thank the Institute of Geology and Geophysics named after H.M. Abdullaev for access to its thin-section collections and to the petrographic descriptions that served as reference data for this evaluation.

Funding

The authors did not receive support from any organization for the submitted work. No funding was received for conducting this study.

Competing interests

The authors have no relevant financial or non-financial interests to disclose.

Separating method failure from reference-data limits in automated thin-section petrography: palaeovolcanic rocks of southern Uzbekistan

Abstract

Automated image analysis promises to remove the main bottleneck of regional petrography, the hours a specialist spends on each thin section, and modern pipelines report high agreement with petrographers on fresh, coarse-grained rocks. Reports of failure on altered material are becoming equally common, but the literature rarely separates two very different causes: the method is genuinely poor, or the reference descriptions used for scoring are too coarse to reward a correct detector. We show that Cohen's κ conflates these causes, because it is bounded above by a ceiling κ^{max} fixed by class prevalence and reference quality rather than by the method. We derive that ceiling, define an evaluability criterion from it, and apply both to Petrolog-AI, a rule-based pipeline tested on 209 image pairs from 96 samples of the Qorason palaeovolcano (South Tien Shan, Uzbekistan). Pre-processing transferred across volcanic centres and reduced the measurement step for the whole collection to about 14 minutes. Identification did not: κ ranged from −0.053 to 0.073, with bootstrap intervals spanning zero. Of 16 group–dataset combinations only 10 are evaluable, and among those the pipeline realises 6 % of its achievable agreement. Prevalence-adjusted coefficients proposed as remedies for the kappa paradox prove unsafe here, scoring an image-free baseline at 0.71–0.95. The negative result is a property of colour-based rules on fine-grained altered rocks, not of the reference data, and the route to higher performance runs through point-level reference databases rather than re-tuned thresholds.

Keywords: thin-section petrography; image analysis; Cohen's κ; evaluation methodology; reference databases; southern Uzbekistan

1. Introduction

Thin-section petrography remains the primary source of mineralogical and textural data in igneous petrology, and it remains slow. A complete description of a single sample by a trained specialist takes several hours (Liu et al. 2022), and manual tracing of grain boundaries has been reported at about 155 minutes per image against roughly 3 minutes with open segmentation software (Azzam et al. 2024). For a regional survey holding thousands of archived sections, that difference is the difference between a dataset that can be quantified and one that cannot. The incentive to automate is therefore not novelty but throughput.

The field has responded. Pixel classifiers using colour in plane- and cross-polarised light reported 81–98 % accuracy for five minerals on unseen samples (Baykan and Yılmaz 2010); colour and texture features with support vector machines reached 99 % on 42 minerals (Naseri and Rezaei Nasab 2023); convolutional encoder–decoder networks segment arenite sections and reproduce 350-point counts in modal composition (Tang et al. 2020). Complete tools now exist for grain segmentation (Azzam et al. 2024), for automatic rock description from image pairs (Paucar et al. 2025), for reservoir rocks at industrial scale (Liu et al. 2022), and for ore minerals in reflected light (Jiang et al. 2024; Acevedo Zamora and Kamber 2023).

Two conditions underlie those results, and both concern data rather than algorithms. The images are controlled scans rather than compressed archive photographs, and the reference is a point count or a pixel-level annotation rather than a written description. Where either condition fails, reported performance falls sharply: Faria et al. (2025) obtained 65 % accuracy on carbonate sections from wells not seen during training, and reviews of the field name the absence of independent benchmarks as its central weakness (Long et al. 2022; Zhao et al. 2024). A negative result on archive material is thus unsurprising. What is missing is a way to read it.

Two structurally different explanations produce the same low κ. Under the first, the method genuinely fails: the colour rules or thresholds do not capture the optical properties of the target assemblage, and a better pipeline would score higher. Under the second, the reference fails: the written description does not reliably record every mineral group, so even a perfect detector would disagree with it, and no pipeline could score higher until the reference improves. The two call for opposite investments — algorithm development in the first case, reference-database construction in the second — and Cohen's κ does not distinguish them.

This paper supplies the missing distinction. We show that κ carries an upper bound κ^{max} determined by the true prevalence of a class and by the sensitivity and specificity of the reference labelling process, derive that bound in closed form, and use it to classify each evaluation case as evaluable or non-evaluable. We then apply the diagnostic to a complete held-out validation of Petrolog-AI, a rule-based pipeline for archive thin-section photographs, on altered felsic volcanic rocks of the Qorason palaeovolcano, southern Uzbekistan. The analysis yields four results: 6 of 16 group–dataset combinations cannot be judged by κ at all; the remaining 10 attain on average 6 % of their achievable agreement, which establishes the failure as genuine; the prevalence-adjusted coefficients usually recommended as a remedy for the kappa paradox would have declared the pipeline successful on precisely the groups where it has no skill; and an intuitive reading of the development-to-test comparison as overfitting is an artefact of prevalence. The contribution is the diagnostic and the evaluation protocol built on it, illustrated by a case in which it reverses the conclusion one would otherwise draw.

2. Materials and methods

2.1. Rock material and image collections

The test material consists of archive thin-section photographs from the Qorason palaeovolcano in the south-western Gissar Range (South Tien Shan, Uzbekistan), which comprises two caldera-type synclines separated by an uplifted anticlinal block (Shoyoqubov 1978). Its felsic volcanics — rhyolite lavas, clastolavas and crystal-clastic tuffs with subordinate dacites and andesites — belong to the Lower Carboniferous (Viséan) Vakhshivar Formation (Bensh 1975) and were altered along faults by silicification, sericitisation, chloritisation and carbonatisation accompanied by pyrite–polymetallic mineralisation. The diagnostic minerals of that overprint — sericite, chlorite and carbonate — are micrometre-scale and intergrown with a cryptocrystalline groundmass, which makes the material a deliberately demanding target for colour-based image analysis and a realistic one for regional survey archives. Sample locations are shown in Fig. 1.

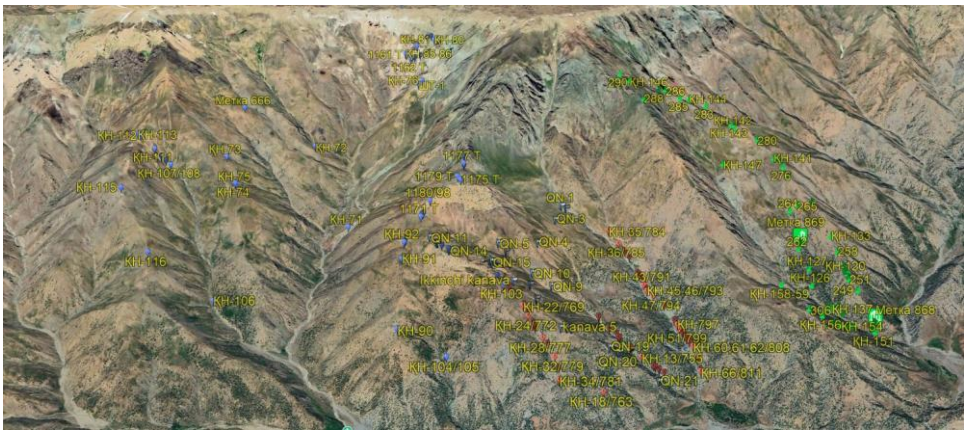

Fig. 1 Sample locations on a satellite view of the Qorason palaeovolcano. KH: archive catalogue samples; QN: newly prepared sections (labelled KN in the image files); numbered and "T" marks are other field stations. Imagery © Google Earth

Two collections are used (Table 1). The development set (KM and BX series) comprises 420 plane-polarised / cross-polarised (PPL/XPL) pairs from 184 samples of the neighbouring Qizilmachit and Vakhshivar palaeovolcanoes, and was used to fix every pipeline threshold. The test set (KH and KN series, Qorason) consists of 209 pairs from 96 samples: 185 pairs from 86 archive catalogue samples, stored as small compressed photographs, and 24 pairs from 10 newly prepared sections imaged at full camera resolution. Because development and test material come from different volcanic centres, the design measures transfer between centres of one region rather than re-substitution. The test images were processed once, after all parameters had been frozen.

Table 1. Image collections.

Collection	Series	Samples	PPL/XPL pairs	Image size (px)	Reference description
Development	KM, BX	184	420	varies	Used to fix thresholds only

Collection	Series	Samples	PPL/XPL pairs	Image size (px)	Reference description
Test, archive	KH	86	185	c. 300–490 × 350–640	Catalogue text: rock name, minerals, structure, texture
Test, new sections	KN	10	24	2464 × 2056	Petrographic report of the same sections
Test, total	–	**96**	**209**	–	

KH images are photographs cropped from a printed catalogue; KN images are full-resolution camera frames. Both carry linear scale bars. The contrast in resolution is itself a test of how the pipeline degrades on archive material.

2.2. Algorithm: image pre-processing and mineral-group rules

Petrolog-AI treats a thin-section photograph as a measurement problem rather than a classification task: every term in its output carries the numeric quantity that produced it, so any statement in the final report can be traced back to a measurement. One image pair passes through seven steps arranged as a closed loop (Fig. 2), of which steps 2–6 are automatic and step 7 returns the result to the petrographer.

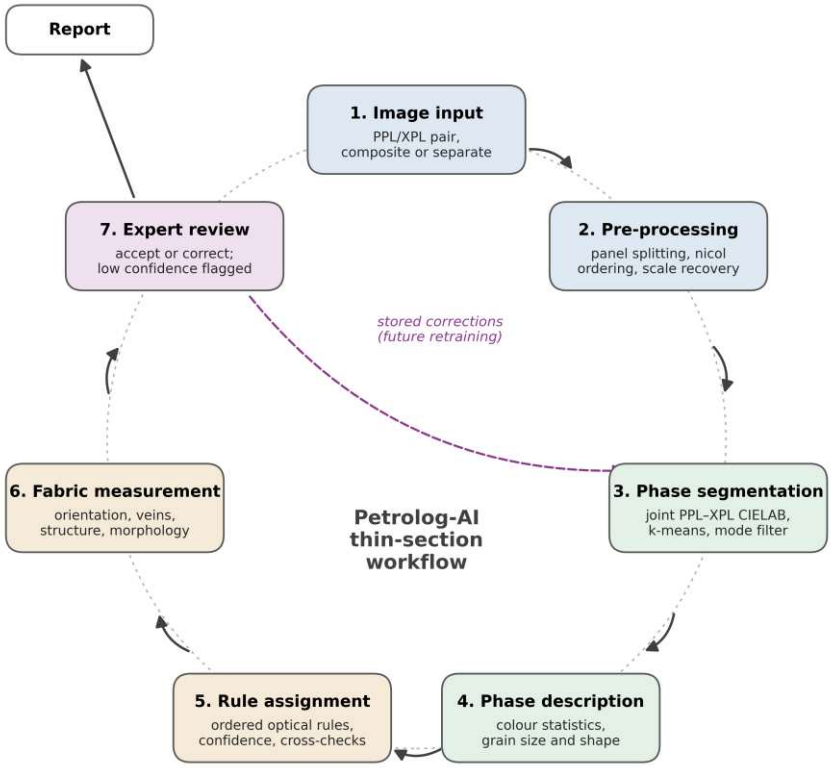

Fig. 2 Petrolog-AI thin-section workflow. Solid arrows show the processing sequence for one image pair; the dashed arrow marks expert corrections, which are stored with the sample for future retraining of the segmentation step but were not used in the version evaluated here

Composite archive images are first split at a seam detected as a column of nearly constant colour or, where two panels touch directly, at the column of largest combined jump in brightness and saturation ($|\Delta V| + 1.5|\Delta S| > 0.18$). The panel with the higher score $2S - V$ in HSV space is taken as XPL, because interference colours raise saturation while grains at extinction lower brightness. Physical scale is recovered from the printed scale bar: candidate bars are sought in the lower-left corner, then along the lower strip and finally in the upper-left corner, their length is measured in pixels, and the label is read by OCR under several binarisations, the modal reading being accepted and rounded to the nearest standard length.

Phases are segmented by k-means clustering ($k = 7$, or 6 when only a PPL image exists) on the six-component vector that joins the CIELAB coordinates of each pixel in both nicols. Joining the nicols is the central

design choice: quartz, feldspar and carbonate are nearly indistinguishable in PPL but separate by XPL brightness and interference colour, while the XPL colour of a single grain varies with stage rotation and is partly stabilised by the PPL component. The label map is then cleaned by a mode filter whose window grows with image size. Each resulting phase is described by its median hue, saturation and value in both nicols, its area fraction, and an area-weighted median grain diameter obtained from the connected components of its mask.

Mineral groups are assigned by an ordered sequence of optical rules applied to those medians, the first satisfied rule deciding the outcome. A phase dark in both nicols ($V_p < 0.20$, $V_x < 0.20$) is assigned to opaque minerals (OPQ); one bright in PPL but dark in XPL to the isotropic group (ISO), which covers volcanic glass and minerals at extinction; one with strong, bright XPL colour ($S_x > 0.30$, $V_x > 0.40$) to high-birefringence minerals (HIB: carbonate, white mica, epidote); one with its own PPL colour at hue 10–90° to coloured mafic minerals (MAF: biotite, chlorite, amphibole); and colourless phases ($S_p < 0.15$) to quartz–feldspar (QF). The rules stop deliberately at group level, because relief, extinction angle, twinning and pleochroism — the properties that separate quartz from untwinned feldspar or biotite from hornblende — are not recorded in a single image pair. Fabric terms follow from measurements: preferred orientation from the structure tensor, porphyritic structure from the grain-size ratio between the coarsest and finest phase, and veining from the elongation and bounding-box fill of opaque components. Fig. 3 shows the segmentation and the resulting assignment for one sample.

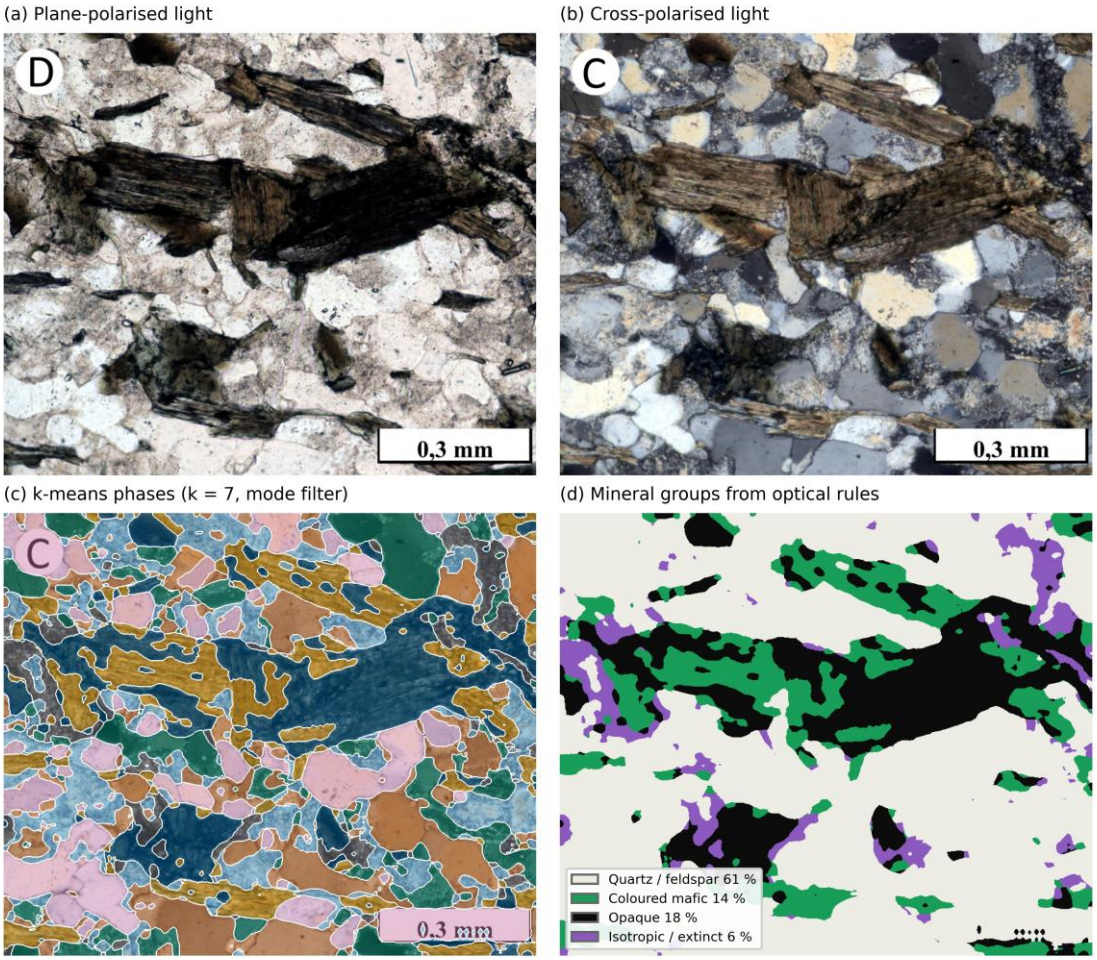

Fig. 3 Phase segmentation of sample QN-15 (Qorason palaeovolcano, objective 10×). (a) Plane-polarised light; (b) cross-polarised light; (c) seven k-means phases from joint PPL–XPL CIELAB colour after mode filtering, overlaid on the XPL image with phase boundaries in white; (d) mineral groups assigned by the ordered optical rules. Scale bar 0.3 mm

2.3. Implementation: sample-level decisions and software environment

The pipeline operates on individual image pairs, whereas the reference descriptions refer to whole sections. A mineral group was therefore counted as present in a sample when its mean modal proportion over all pairs of

that sample reached a group-specific threshold: 40 % for QF, 10 % for MAF, 1 % for HIB, 2 % for OPQ and 10 % for ISO. The thresholds were chosen on the development set by maximising Cohen's κ over a small grid. A fabric feature was counted as present when the median measured value over the sample's pairs exceeded its limit. No threshold was adjusted after the test set had been processed.

The pipeline is implemented in Python 3.10 and uses NumPy 1.24, OpenCV 4.8, Matplotlib 3.7 and Tesseract OCR 4.1. It runs on a single CPU core of a standard desktop computer and requires no graphics accelerator. The program contains no trained statistical model: every identification step is a fixed, inspectable rule, so any sample-level decision can be traced back to the measurement and threshold that produced it. A large language model (Claude, Anthropic) was used as a programming aid while the evaluation scripts were written and while the numerical results reported in Section 4 were cross-checked. Every quantity reported in this paper was recomputed from the primary data and verified by the authors, who take full responsibility for the content.

2.4. Testing: reference data

Reference labels were derived from written petrographic descriptions prepared independently of this study. Because those descriptions are qualitative, the comparison is made at the level the text supports. Minerals named in a description were mapped to the five groups: quartz, albite, plagioclase and K-feldspar to QF; biotite, chlorite, amphibole and pyroxene to MAF; sericite, muscovite, carbonate, epidote and titanite to HIB; sulphides, pyrite, magnetite and hematite to OPQ; volcanic glass, garnet and fluorite to ISO. Porphyritic structure, oriented texture and veining were extracted by keyword. All 96 test samples and all 184 development samples carry a description sufficient for both the mineral-group and the fabric comparison, so every sample enters every contrast reported below.

This mapping is the weak link of the evaluation, and the reason the present analysis is necessary. A description that does not mention chlorite is not evidence that chlorite is absent; it may be evidence that the describing petrographer did not consider it worth recording. Section 3 makes the consequences of that ambiguity quantitative instead of leaving them as a caveat.

2.5. Testing: evaluation metrics and baseline

Agreement was scored at sample level by Cohen's κ (Cohen 1960) together with precision, recall and F1. κ is emphasised because several groups are strongly imbalanced and raw accuracy would reward a system that always predicts the majority class. Uncertainty on κ is given as a bias-corrected bootstrap interval over 5000 resamples of the 96 samples (Efron and Tibshirani 1993). As a reference point, an image-free frequency baseline predicts every group whose development-set prevalence exceeds 50 %; in practice it predicts QF for every sample and nothing else. Any claim that the pipeline extracts information from the images must survive comparison with that baseline.

3. The κ^{max} diagnostic

3.1. The ceiling on Cohen's κ under an imperfect reference

Consider a binary task in which the true prevalence of the positive class is π and the reference labels are produced by an imperfect observer with sensitivity s (probability of labelling a true positive as positive) and specificity t (probability of labelling a true negative as negative). Such a reference marks positives at a rate

$$p^{+} = \pi s + (1 - \pi)(1 - t), \qquad (1)$$

which differs from π whenever the observer errs. A perfect detector — one whose call always equals the true label — nonetheless disagrees with this reference on every item the observer mislabelled, so its observed agreement is

$$p_o = \pi s + (1 - \pi)t, \qquad (2)$$

the largest observed agreement attainable by any detector against this reference. The chance-agreement term formed from the same two marginals — the detector's positive rate π and the reference's positive rate p^{+} — is

$$p_e = \pi\, p^{+} + (1 - \pi)(1 - p^{+}), \qquad (3)$$

and the ceiling on κ follows directly from its definition:

$$\kappa^{max} = (p_o - p_e) / (1 - p_e). \qquad (4)$$

Equation (4) depends only on π, s and t — on the structure of the problem and the quality of the reference — and not on the detector being evaluated. When $s = t = 1$ it returns 1 for every π, as it must. As either falls, κ^{max} falls with it, and the decline is far steeper at extreme prevalence than near $\pi = 0.5$. The reason is structural: at $\pi = 0.05$ the reference contains about five true positives in a hundred items, so a sensitivity of 0.7 removes one and a specificity of 0.9 adds about ten false ones, and the positive margin that κ divides by is dominated by noise. This is the mechanism behind the kappa paradox of Feinstein and Cicchetti (1990) and Byrt et al. (1993), expressed as a bound rather than as an anomaly.

3.2. Estimation and the evaluability criterion

In practice π is unknown, but p^+, the reference's own positive rate, is observed directly. Inverting Eq. (1) gives $\hat{\pi} = (p^+ - (1 - t)) / (s - (1 - t))$, clipped to [0.001, 0.999]. Two reference-quality scenarios are carried through the analysis: a moderate scenario ($s = 0.80$, $t = 0.90$), intended to represent a competent petrographer writing free text in which a mineral present in small amount is sometimes left unmentioned, and a conservative scenario ($s = 0.70$, $t = 0.85$) for descriptions of lower or uneven quality. Neither is estimated from the data; both are stated assumptions whose influence is reported explicitly.

A group–dataset combination is called *evaluable* when $\kappa^{max} \geq 0.30$ under the moderate scenario, so that a competent classifier could in principle reach at least fair agreement on the conventional scale of Landis and Koch (1977). Where $\kappa^{max} < 0.30$ the case is *non-evaluable*: a low observed κ is then compatible with any level of method quality, and reporting it as evidence of failure is unsound. The threshold is a convention, not a property of the data, and the analysis below reports how the classification changes when it is applied under the conservative scenario instead.

3.3. Why not a prevalence-adjusted coefficient?

An established response to the kappa paradox is to replace κ with a coefficient whose chance term does not depend on the observed marginals. The prevalence-adjusted bias-adjusted kappa, $PABAK = 2p_o - 1$ (Byrt et al. 1993), and Gwet's AC1 (Gwet 2008) are the two in common use. Both are well motivated as descriptions of agreement, and both are unsuitable as evidence of method skill in the setting considered here, for a reason that Section 4.3 demonstrates empirically: by construction they reward high raw agreement, and at extreme prevalence high raw agreement is available without using the data at all. A classifier that never predicts a rare class agrees with the reference on almost every sample and is scored accordingly.

The ceiling takes the opposite approach. It leaves κ unchanged — retaining the property that a non-informative classifier scores zero — and asks instead whether the observed κ carries interpretable information. Replacing the metric answers the question "how much do the two raters agree?"; computing κ^{max} answers "can this comparison distinguish a good detector from a bad one?". Only the second question bears on whether a method should be developed further or abandoned.

3.4. Field-to-field stability as an internal consistency check

The pipeline is deterministic given a fixed image, so variation in its calls across different image pairs from one sample reflects genuine optical variation — stage position, magnification, local fabric — rather than algorithmic noise. High within-sample stability indicates that aggregating pairs to a sample-level decision is meaningful; low stability indicates that the call fluctuates with field position, in which case a whole-section reference description is summarising a mixture and the sample-level comparison is correspondingly blunt. Stability is measured as the proportion of field–field pairs within a sample for which both fields return the same binary call, under the thresholds of Section 2.3. The measure requires no reference data and is therefore available for any collection of multi-image samples.

4. Results

4.1. Pre-processing: throughput and the pipeline's own control

Pre-processing serves two purposes in this evaluation. It is the step with the clearest practical payoff, and it is the only step whose output can be checked without a petrographic description: a recovered scale is either consistent with the printed bar or not, and nicol order is known independently for the KN pairs. It therefore functions as a positive control for the whole analysis.

A physical scale was recovered for 418 of 420 development pairs (99.5 %) and for 132 of 185 Qorason archive pairs (71 %), in 116 of the latter with at least two concordant OCR readings. None of the scale bars on the 24 new KN photographs was found: they are drawn in a style absent from the archive material on which the detector was developed, which is a coverage gap of the detector rather than a failure of character recognition. Nicol order was identified correctly for 20 of the 24 KN pairs with known labels (83 %). About one third of the image area was routed to the expert as low-confidence in every set, consistent with the design principle that an automated petrographic tool should show where it is uncertain (Petrelli 2024; Ágreda-López and Petrelli 2025).

The time budget is the result with the most direct bearing on regional practice. Median processing time was 2.3 s per pair on the development set and 3.9 s on the archive test set, so the measurement step for the entire Qorason collection of 209 pairs occupied about 14 minutes on a single core of an ordinary office computer. The 24 full-resolution KN frames took about 40 s each, almost entirely because the scale-bar search runs before the image is reduced — an implementation cost, not an algorithmic one. These figures are not directly comparable with the several hours a specialist needs for a complete description (Liu et al. 2022), because the pipeline does not produce a complete description; what they establish is that the automated fraction of the work adds a negligible time cost, so the economic case for automation rests entirely on how much of the description it can be made to carry. That is the question the rest of this paper addresses.

4.2. Agreement on mineral groups

Agreement between the pipeline and the reference descriptions on mineral-group presence was close to chance throughout (Table 2). Every bootstrap interval includes zero. QF reached an F1 of 0.75 and a recall of 1.00 with $\kappa = 0.000$: the group is named in 57 of the 96 samples and the pipeline reports it in all 96, giving 39 false positives, no false negatives, and a marginal distribution indistinguishable from chance. Its bootstrap interval is degenerate at [0.000, 0.000] because the pipeline never varies. MAF was recovered in 10 % of the samples in which it is described and HIB in 5 % — these are the groups holding the chlorite, sericite and carbonate that record alteration. OPQ showed the highest point estimate ($\kappa = 0.073$) and ISO was over-reported eightfold.

Table 2. Sample-level detection of mineral groups on the test set (96 samples). κ^{max} is computed at $s = 0.80$, $t = 0.90$.

Group	Ref. prev.	TP	FP	FN	FP/FN	F1	κ	95 % CI	κ^{max}
QF	59.4 %	57	39	0	∞	0.75	0.000	[0.000, 0.000]	0.63
MAF	32.3 %	3	3	28	0.11	0.16	0.064	[−0.075, 0.223]	0.70
HIB	42.7 %	2	3	39	0.08	0.09	−0.006	[−0.103, 0.096]	0.70
OPQ	22.9 %	12	33	10	3.30	0.36	0.073	[−0.100, 0.243]	0.64
ISO	5.2 %	1	39	4	9.75	0.04	−0.053	[−0.144, 0.039]	**0.01**[a]

Ref. prev. is the proportion of the 96 samples in which the group is named in the reference description. Intervals are bias-corrected bootstrap percentiles over 5000 resamples. Superscript a marks a non-evaluable case (ceiling below 0.30). The QF interval is degenerate because the pipeline predicts the group for every sample.

4.3. Prevalence-adjusted coefficients reward the uninformative classifier

Table 3 scores the same comparisons with PABAK and Gwet's AC1 alongside κ, and adds the image-free baseline of Section 2.5 as a control. The result is unambiguous. κ assigns the baseline exactly zero on every group, which is the correct answer for a classifier that reads no images. PABAK assigns it between 0.19 and 0.90, and AC1 between 0.40 and 0.95. On ISO, the rarest group, AC1 rates the baseline at 0.945 — "almost perfect

agreement" on the conventional scale — for the single rule "never report isotropic phases". On OPQ it rates the baseline at 0.712, or "substantial".

Table 3. Three agreement coefficients applied to the pipeline and to an image-free baseline (test set, 96 samples).

	Pipeline			Image-free baseline		
Group	**κ**	**PABAK**	**AC1**	**κ**	**PABAK**	**AC1**
QF	0.000	0.188	0.399	0.000	0.188	0.399
MAF	0.064	0.354	0.531	0.000	0.354	0.557
HIB	−0.006	0.125	0.312	0.000	0.146	0.357
OPQ	0.073	0.104	0.179	0.000	**0.542**	**0.712**
ISO	−0.053	0.104	0.301	0.000	**0.896**	**0.945**

The baseline predicts quartz–feldspar for every sample and no other group, and uses no image information whatsoever. PABAK = $2p_o - 1$ (Byrt et al. 1993); AC1 after Gwet (2008). Bold values mark cases where a coefficient rates the uninformative baseline as substantial or almost perfect agreement.

Two consequences follow. First, had this study replaced κ with either coefficient in response to the kappa paradox, it would have reported substantial to almost perfect agreement for the two groups on which the pipeline demonstrably has no skill, and would have concluded that the method works. Second, on OPQ and ISO the pipeline scores lower than the baseline under both adjusted coefficients, because it over-reports those groups while the baseline stays silent; a reader comparing adjusted coefficients alone would conclude that using the images is worse than ignoring them, which is true of raw agreement and uninformative about mineralogy. The problem that PABAK and AC1 are designed to solve is real, but at extreme prevalence they solve it by making the metric insensitive to exactly the signal under test.

4.4. Evaluability under κ^{max}

Table 4 reports κ and κ^{max} for all 16 group–dataset combinations — five mineral groups and three fabric features, on each of the two datasets. Ten are evaluable under the moderate scenario. The six that are not comprise test ISO (prevalence 5.2 %), test oriented texture (11.5 %), development OPQ (11.4 %), development ISO (12.0 %), development oriented texture (8.7 %) and development vein texture (3.8 %). In every case the cause is the same: a class too rare for the chance-correction term to operate as intended.

Table 4. κ, κ^{max} and their ratio for all 16 group–dataset combinations.

	Test set (n = 96)			Development set (n = 184)		
Group or feature	**κ**	**κ^{max}**	**κ/κ^{max}**	**κ**	**κ^{max}**	**κ/κ^{max}**
Quartz–feldspar (QF)	0.000	0.63	0.00	−0.006	0.70	−0.01
Coloured mafic (MAF)	0.064	0.70	0.09	−0.019	0.49	−0.04
High-birefringence (HIB)	−0.006	0.70	−0.01	0.124	0.60	0.21
Opaque (OPQ)	0.073	0.64	0.11	0.114	**0.21**[a]	—
Isotropic (ISO)	−0.053	**0.01**[a]	—	0.135	**0.27**[a]	—
Porphyritic structure	0.067	0.66	0.10	0.160	0.62	0.26
Oriented texture	−0.018	**0.22**[a]	—	0.026	**0.01**[a]	—
Vein texture	−0.049	0.45	−0.11	0.080	**0.01**[a]	—
Mean over the 10 evaluable cases	**0.041**	**0.619**		**0.061**		

Superscript a: non-evaluable — ceiling below 0.30 under the moderate scenario. All 96 test and 184 development samples enter every contrast. Ratios are reported only for evaluable cases, since dividing by a ceiling near zero is not meaningful.

4.5. Agreement among the evaluable cases

Across the ten evaluable cases the mean κ is 0.041 against a mean ceiling of 0.619, a mean ratio of 0.061: the pipeline realises about 6 % of the agreement available to it under the stated reference quality. Individual ratios range from −0.11 to 0.26, the highest being porphyritic structure on the development set (κ = 0.160, ceiling 0.62).

No evaluable case reaches $\kappa > 0.20$. Under the conservative scenario the picture does not improve but the accounting changes: development MAF (ceiling 0.09) and test vein texture (ceiling 0.01) drop below the criterion, leaving eight evaluable cases with a mean ratio of 0.143. The ratio rises because the ceiling falls, not because agreement improves, and the conclusion that the pipeline captures a small fraction of the achievable signal holds under both assumptions.

4.6. Error structure

The errors are strongly asymmetric, and the direction differs by group in a way that identifies the underlying optical cause. QF produces 39 false positives and no false negatives: colourless phases of moderate XPL brightness occur somewhere in essentially every image, so the rule fires universally. MAF and HIB reverse this — 3 false positives against 28 and 39 false negatives — because the sericite, chlorite and carbonate of the alteration assemblage are micrometre-scale, intergrown with the groundmass, and colour-ambiguous at a single stage position, so the rules miss them systematically rather than randomly. OPQ shows the opposite asymmetry (33 false positives against 10, FP/FN = 3.3), consistent with dark groundmass pixels satisfying the darkness criterion in both nicols. ISO is over-reported ninefold for the same reason. The pattern is not noise: it is what colour rules designed for millimetre-scale, well-separated grains produce when applied to a cryptocrystalline matrix.

Aggregation method matters for one group only. Replacing the mean modal proportion by an any-field rule raises OPQ κ from 0.073 to 0.214, which indicates that opaque grains are genuinely present but not consistently above the per-pair threshold in every image of a sample. For QF, MAF, HIB and ISO the choice of mean, any, majority or all changes κ by less than 0.1, so the sample-level decision rule is not the limiting factor.

4.7. Field-to-field stability and the development-to-test comparison

Within-sample field-to-field agreement (Table 5) is high for QF (93.9 %) and HIB (95.4 %) but much lower for OPQ (64.9 %) and ISO (72.5 %); for OPQ more than a third of multi-image samples contain a field that disagrees with the others. Because the pipeline is deterministic, this instability measures real variation between fields of the same section, not algorithmic noise. It carries a practical implication: for OPQ and ISO a whole-section reference label is summarising a heterogeneous section, so the sample-level comparison is intrinsically blunt for precisely the two groups whose ceilings are lowest.

Table 5. Within-sample field-to-field stability (test set: 96 samples, 209 pairs, 131 field–field comparisons).

Group	Field–field agreement	Samples with ≥ 1 discordant field
QF	93.9 %	7.3 %
MAF	90.8 %	9.4 %
HIB	95.4 %	5.2 %
OPQ	64.9 %	36.5 %
ISO	72.5 %	32.3 %

Computed over all pairs of image fields within each multi-image sample, using the sample-level thresholds of Section 2.3.

The development set returned $\kappa = 0.114$ for OPQ and $\kappa = 0.135$ for ISO, the two highest mineral-group values on either dataset, and neither reproduced on the test set. The natural reading of that pattern is modest agreement acquired on the data used to fix the thresholds and then lost on held-out material — overfitting. The ceiling refutes it. Development OPQ has $\kappa^{max} = 0.21$ at a prevalence of 11.4 %, and development ISO has $\kappa^{max} = 0.27$ at 12.0 %; both are non-evaluable. The apparent agreement was never evidence of skill. It arose because low prevalence compresses κ and its ceiling toward zero together, so a small numerator occupies a large share of a small range. The drop between datasets tracks the change in prevalence — OPQ rises from 11.4 % to 22.9 %, ISO falls from 12.0 % to 5.2 % — rather than any change in method quality. Without the ceiling, this study would have reported overfitting and recommended regularisation; with it, the correct recommendation is to obtain reference data in which these groups are recorded systematically.

5. Discussion

5.1. The failure of colour-based identification is genuine

For the ten evaluable cases the diagnostic confirms the negative result rather than explaining it away. A mean κ/κ^{max} of 0.061 means the pipeline recovers about one fifteenth of the agreement a perfect detector would obtain against the same imperfect reference. This cannot be attributed to the reference: the moderate scenario already allows one true positive in five to go unrecorded and one true negative in ten to be recorded spuriously, and the conclusion survives the conservative scenario as well. The responsibility lies with the method.

The mechanism is identifiable from the error structure of Section 4.6 and is specific to the material. Colour-based rules assume that a mineral occupies a compact region of colour space and that grains are large enough for their colour to be measured. In hydrothermally altered rhyolite neither assumption holds. Sericite and chlorite form micrometre-scale intergrowths whose apparent colour is a mixture of several phases within one pixel; carbonate shares its high-birefringence signature with white mica; and the XPL colour of any anisotropic grain depends on its orientation relative to the polarisers, so a single image pair samples one arbitrary point of a continuum. The groups that failed most completely, MAF and HIB, are exactly those containing the alteration minerals; the group that failed in the opposite direction, QF, is the one whose optical signature — absence of colour — is shared by the groundmass itself. These are limits of the image and the rule class, not of threshold tuning, and no adjustment of the five thresholds in Section 2.3 would alter them.

5.2. Non-evaluable cases require different data, not a better algorithm

Six combinations cannot be judged by κ. This is not a statement that the pipeline fails on them; it is a statement that the comparison carries no information either way. The distinction has immediate practical force. If a future version raised test OPQ κ from 0.073 to 0.20 against a ceiling of 0.64, that would be real progress and could be reported as such. An equivalent movement in test ISO κ could not, because with a ceiling of 0.01 the metric cannot separate a good detector from a bad one; improvement and deterioration are both within noise. Reporting a mean κ pooled across evaluable and non-evaluable classes, as is common, mixes measurements with non-measurements.

5.3. What a stronger reference database would change

The analysis identifies the binding constraint on automated petrography in a regional archive setting, and it is not computational. Every non-evaluable case in Table 4 arises from a class whose prevalence in the reference is too low, and prevalence in the reference is a property of how descriptions were written, not of the rocks. A point-counted reference, or even a structured checklist in which each of the five groups is explicitly recorded as present or absent, would raise reference sensitivity and specificity and would record rare groups systematically rather than when they happened to catch the describer's attention. Both effects raise the ceiling, and they do so without any change to the pipeline.

The magnitude is worth stating. At the observed prevalences, moving reference quality from the conservative to the moderate scenario raises the mean ceiling over the eight cases evaluable under both from 0.464 to 0.656. A reference approaching $s = t = 0.95$ would lift 14 of the 16 cases above the evaluability criterion. The two that remain below it are instructive. Test ISO and development vein texture are named in only 5.2 % and 3.8 % of their samples, and at that rate the inversion of Eq. (1) attributes nearly the whole observed positive margin to reference false positives: the inferred true prevalence approaches zero and the ceiling collapses however good the reference becomes. For classes that rare the remedy is not a better description of the same sections but a sample set in which the class actually occurs, which is a sampling decision rather than a labelling one. The same reasoning applies to supervised learning, which this pipeline deliberately avoids: a convolutional classifier trained on these descriptions would inherit their noise as a ceiling on its training signal as well as on its evaluation. The literature's high reported accuracies (Tang et al. 2020; Naseri and Rezaei Nasab 2023) rest on point counts and pixel annotations, and the gap between those results and the present one is in substantial part a gap in reference infrastructure rather than in algorithms. For a survey archive, building that infrastructure for a few hundred sections is a larger investment than running a pipeline, and a more durable one: the labels outlive any particular model.

The division of labour that emerges is consistent with the view that machine learning in petrology should automate routine measurement and leave interpretation with the specialist (Petrelli 2024; Ágreda-López and Petrelli 2025). On present evidence the boundary for altered felsic volcanics falls between measurement, which transfers across centres at about 4 s per image pair, and identification, which does not transfer at all. That boundary is not fixed; it is where the available reference data place it.

5.4. Recommendations for validation practice

Three practices follow, none of which requires new experiments. First, compute κ^{max} for every class before reporting κ. The computation needs only the reference prevalence and defensible bounds on reference sensitivity and specificity, which follow from the type of reference used — point count, structured checklist, free-text description, keyword extraction — and should be stated as assumptions and varied, as here. Second, classify each case as evaluable or non-evaluable and report the two groups separately; pooling them into a single mean conceals both genuine failures and spurious successes. Third, record field-to-field stability wherever samples carry several images: it requires no reference data at all and reveals when a whole-sample label is summarising a heterogeneous section. To these we add a caution drawn from Section 4.3: a prevalence-adjusted coefficient is not a safe substitute for κ when the question is whether a method extracts information, because at extreme prevalence such coefficients reward silence.

5.5. Limitations

The ceiling model assumes that reference errors are independent of the pipeline's predictions and that reference sensitivity and specificity are constant across samples. Both are idealisations. If the cataloguing petrographer applied different standards to different rock types, or if the pipeline's errors correlate with the conditions under which the describer erred — both plausible where fine grain size impedes human and machine alike — then κ^{max} is misestimated, and most likely overestimated, which would strengthen rather than weaken the conclusion that the method underperforms. Reporting two scenarios bounds the sensitivity of the result to these assumptions but does not eliminate them. Estimating s and t directly would require a subset of sections described twice, independently, which this archive does not contain. Beyond the model, the study covers one palaeovolcano, two collections and presence labels derived from text; modal proportions, grain-size accuracy, the polished-section module and the geochemical module were not evaluated, and the conclusions extend only to colour-rule identification of mineral groups in altered felsic volcanic rocks.

6. Conclusions

1. Cohen's κ carries an upper bound set by the true prevalence of a class and by the sensitivity and specificity of the reference labelling process. Equation (4) gives that bound in closed form; it depends on the structure of the evaluation and not on the method under test, and comparing an observed κ with it separates genuine method failure from the limits of the reference data.

2. Applied to a held-out validation of Petrolog-AI on 209 image pairs from 96 Qorason samples, the diagnostic classifies 6 of 16 group–dataset combinations as non-evaluable, because mineral prevalence in the reference descriptions suppresses the ceiling below 0.30. Among the 10 evaluable cases the mean κ/κ^{max} is 0.061, which establishes the near-chance agreement as genuine method failure rather than a deficiency of the reference.

3. Prevalence-adjusted coefficients are not a safe substitute. On this dataset PABAK and Gwet's AC1 score an image-free baseline between 0.19 and 0.95, rating it "almost perfect" on the rarest group, whereas κ correctly returns zero. Any study that adopts such a coefficient in response to the kappa paradox should verify it against an uninformative baseline first.

4. The apparent development-set agreement for opaque and isotropic phases, which would ordinarily be read as overfitting lost on held-out data, is an artefact of low prevalence in those datasets. The diagnostic reverses the diagnosis and, with it, the remedy: the problem is reference data, not regularisation.

5. Automation delivers unambiguously where the measurement is objective and verifiable. Scale was recovered for 99.5 % of development pairs and 71 % of archive pairs, nicols were ordered correctly for 83 % of labelled pairs, and the whole 209-pair collection was processed in about 14 minutes on one CPU core against the several hours a specialist requires for a single complete description. Colour-based identification of mineral groups does not transfer to altered felsic volcanics, because the diagnostic minerals are micrometre-scale, intergrown and colour-ambiguous at a single stage position.

6. The binding constraint is reference infrastructure rather than computation. Every non-evaluable case traces to a group recorded too rarely in free-text descriptions, and a point-counted or structured reference would raise the ceiling for all sixteen cases without any change to the pipeline. This is the practical lesson for regional petrography: the investment that most increases what automation can deliver is a systematically labelled reference collection, which outlives any particular model.

7. Automated petrographic analysis is at an early stage in Uzbekistan, where regional archives hold large collections of thin sections and compressed catalogue photographs that have never been analysed quantitatively. The results reported here indicate how such collections can be brought into a modern workflow: the measurement steps can be automated immediately and at negligible cost, while identification should remain with the petrographer until point-level reference data, texture-aware classifiers and multi-angle polarised imaging are available. Reporting the evaluability of each comparison alongside its score, as proposed here, is what allows such a programme to distinguish real progress from movement in a metric that cannot measure it.

List of abbreviations

AC1 – Gwet's first-order agreement coefficient
CIELAB – CIE L*a*b* colour space
HIB – high-birefringence minerals (carbonate, white mica, epidote)
HSV – hue–saturation–value colour space
ISO – isotropic phases (volcanic glass, garnet, fluorite)
MAF – coloured mafic minerals (biotite, chlorite, amphibole)
OCR – optical character recognition
OPQ – opaque minerals
PABAK – prevalence-adjusted bias-adjusted kappa
PPL – plane-polarised light
QF – quartz–feldspar
XPL – crossed polarisers
κ – Cohen's kappa coefficient of agreement
κ^{max} – the upper bound on κ imposed by class prevalence and reference quality

Statements and Declarations

Funding

The authors did not receive support from any organization for the submitted work. No funding was received for conducting this study.

Competing interests

The authors have no relevant financial or non-financial interests to disclose. The authors have no competing interests to declare that are relevant to the content of this article.

Ethics approval

Not applicable. This study did not involve human participants, human tissue, animals or personal data, and therefore required no approval by an ethics committee.

Consent to participate

Not applicable. No human participants were involved in this study.

Consent for publication

Not applicable. The manuscript contains no data or images relating to an identifiable individual.

Data availability

The evaluation scripts, per-pair results and reference labels are available at https://github.com/yuldoshevz1995-spec/petrolog-cageo (MIT licence, release v1.0.0). Thin-section photographs and the original petrographic descriptions belong to the collections of the Institute of Geology and Geophysics named after H.M. Abdullaev and are available on reasonable request, subject to institutional permission.

Code availability

Program name: Petrolog-AI, release v1.0.0. Language: Python ≥ 3.10. Dependencies: NumPy ≥ 1.24, OpenCV ≥ 4.8, Matplotlib ≥ 3.7 and Tesseract OCR ≥ 4.1. Hardware: one CPU core of a standard desktop or laptop computer; no GPU required. Repository: https://github.com/yuldoshevz1995-spec/petrolog-cageo (MIT licence).

Author contributions

Zokir D. Yoldoshev: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. Usmonjon D. Mamarozikov: Conceptualization, Supervision, Validation, Writing – review & editing. Ahmadali H. Hamrayev: Investigation, Resources, Writing – review & editing. Nuriddin E. Saitov: Investigation, Resources, Validation. Utkir A. Normatov: Investigation, Data curation. Zilola A. Keldibekova: Data curation, Writing – review & editing. Ulugbek H. Tukhtayev and Utkir O. Ikromov: Investigation, Resources. All authors read and approved the final manuscript.

References

Acevedo Zamora MA, Kamber BS (2023) Petrographic microscopy with ray tracing and segmentation from multi-angle polarisation whole-slide images. Minerals 13:156. https://doi.org/10.3390/min13020156

Ágreda-López M, Petrelli M (2025) Opportunities, epistemological assessment and potential risks of machine learning applications in volcano science. Artificial Intelligence in Geosciences 6:100153. https://doi.org/10.1016/j.aiig.2025.100153

Azzam F, Blaise T, Brigaud B (2024) Automated petrographic image analysis by supervised and unsupervised machine learning methods. Sedimentologika. https://doi.org/10.57035/journals/sdk.2024.e22.1594

Baykan NA, Yılmaz N (2010) Mineral identification using color spaces and artificial neural networks. Computers & Geosciences 36:91–97. https://doi.org/10.1016/j.cageo.2009.04.009

Bensh FR (1975) Stratigrafiya i foraminifery kamennougol¸nykh otlozheniy yugo-zapadnykh otrogov i yuzhnogo sklona Gissarskogo khrebta [Stratigraphy and foraminifera of the Carboniferous deposits of the south-western spurs and southern slope of the Gissar Range]. Tashkent (in Russian)

Byrt T, Bishop J, Carlin JB (1993) Bias, prevalence and kappa. Journal of Clinical Epidemiology 46:423–429. https://doi.org/10.1016/0895-4356(93)90018-V

Cohen J (1960) A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20:37–46. https://doi.org/10.1177/001316446002000104

Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman & Hall, New York. https://doi.org/10.1201/9780429246593

Faria EL, Barbosa R, Coelho JM, Matos TF, Santos BCC, Gonzalez JL, Bom CR, de Albuquerque MP, Russano PJ, de Albuquerque MP (2025) Automatic classification of carbonatic thin sections by computer vision techniques and one-vs-all models. Artificial Intelligence in Geosciences 6:100117. https://doi.org/10.1016/j.aiig.2025.100117

Feinstein AR, Cicchetti DV (1990) High agreement but low kappa: I. The problems of two paradoxes. Journal of Clinical Epidemiology 43:543–549. https://doi.org/10.1016/0895-4356(90)90158-L

Gwet KL (2008) Computing inter-rater reliability and its variance in the presence of high agreement. British Journal of Mathematical and Statistical Psychology 61:29–48. https://doi.org/10.1348/000711006X126600

Jiang C, Abdul Halin A, Yang B, Abdullah LN, Manshor N, Perumal T (2024) Res-UNet ensemble learning for semantic segmentation of mineral optical microscopy images. Minerals 14:1281. https://doi.org/10.3390/min14121281

Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174. https://doi.org/10.2307/2529310

Liu H, Ren Y-L, Li X, Hu Y-X, Wu J-P, Li B, Luo L, Tao Z, Liu X, Liang J, Zhang Y-Y, An X-Y, Fang W-K (2022) Rock thin-section analysis and identification based on artificial intelligent technique. Petroleum Science 19:1605–1621. https://doi.org/10.1016/j.petsci.2022.03.011

Long T, Zhou Z, Hancke G, Bai Y, Gao Q (2022) A review of artificial intelligence technologies in mineral identification: classification and visualization. Journal of Sensor and Actuator Networks 11:50. https://doi.org/10.3390/jsan11030050

Naseri A, Rezaei Nasab A (2023) Automatic identification of minerals in thin sections using image processing. Journal of Ambient Intelligence and Humanized Computing 14:3369–3381. https://doi.org/10.1007/s12652-021-03474-5

Paucar S, Mejia-Escobar C, Collaguazo V (2025) Automatic description of rock thin sections: a web application. Artificial Intelligence in Geosciences 6:100118. https://doi.org/10.1016/j.aiig.2025.100118

Petrelli M (2024) Machine learning in petrology: state-of-the-art and future perspectives. Journal of Petrology 65:egae036. https://doi.org/10.1093/petrology/egae036

Shoyoqubov T (1978) Geologiya i rudonosnost¸ paleozoyskikh vulkanogennykh formatsiy Yuzhnogo Uzbekistana [Geology and ore potential of the Palaeozoic volcanogenic formations of southern Uzbekistan]. Tashkent (in Russian)

Tang DG, Milliken KL, Spikes KT (2020) Machine learning for point counting and segmentation of arenite in thin section. Marine and Petroleum Geology 120:104518. https://doi.org/10.1016/j.marpetgeo.2020.104518

Zhao T, Wang S, Ouyang C, Chen M, Liu C, Zhang J, Yu L, Wang F, Xie Y, Li J et al (2024) Artificial intelligence for geoscience: progress, challenges, and perspectives. The Innovation 5:100691. https://doi.org/10.1016/j.xinn.2024.100691

Figure captions

Fig. 1 Sample locations on a satellite view of the Qorason palaeovolcano. KH: archive catalogue samples; QN: newly prepared sections (labelled KN in the image files); numbered and "T" marks are other field stations. Imagery © Google Earth

Fig. 2 Petrolog-AI thin-section workflow. Solid arrows show the processing sequence for one image pair; the dashed arrow marks expert corrections, which are stored with the sample for future retraining of the segmentation step but were not used in the version evaluated here

Fig. 3 Phase segmentation of sample QN-15 (Qorason palaeovolcano, objective 10×). (a) Plane-polarised light; (b) cross-polarised light; (c) seven k-means phases from joint PPL–XPL CIELAB colour after mode filtering, overlaid on the XPL image with phase boundaries in white; (d) mineral groups assigned by the ordered optical rules. Scale bar 0.3 mm