Check for updates

RESEARCH ARTICLE

REVISED Automated detection of freeze-thaw signatures in archaeological sediments using deep learning

[version 2; peer review: 3 approved, 1 approved with reservations]

Sofia Kouki[1], Li Li[1-4], Vera Aldeias[1]

1Universidade do Algarve, The Interdisciplinary Center for Archaeology and the Evolution of Human Behaviour ICArHEB, Faro, Campus de Gambelas, 8005-139, Portugal
2 Department of Human Origins, Max Planck Institute for Evolutionary Anthropology, Leipzig, 04103, Germany
3 Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences, No. 142 Xizhimenwai Street, Beijing, China, 100044
4 Key Scientific Research Base on Paleolithic Human Evolution and Paleogenetics (IVPP), SACH, Beijing, China, 100044

v2 **First published:** 02 May 2026, **6**:144
https://doi.org/10.12688/openreseurope.23697.1
Latest published: 04 Aug 2026, **6**:144
https://doi.org/10.12688/openreseurope.23697.2

Abstract

Background

Freeze-thaw processes leave diagnostic traces in archaeological soils and sediments that are central to reconstructing past climates and understanding hominin adaptations to glacial environments. Frost features can be defined through soil micromorphology techniques using well-defined diagnostic criteria, but these traces can be subtle, variably expressed, and overlap with other pedogenic processes, making their identification time-consuming, expert-dependent, and subject to high inter-observer variability.

Methods

We trained five convolutional neural network architectures on photomicrographs from eleven Plio-Pleistocene archaeological sites, implementing a two-stage classification approach by detecting first the presence of those features (binary classification) and then the feature type (multiclass classification), and validated model outputs against both interpretability analysis and a blind survey of practicing micromorphologists.

Results

Results reveal a performance paradox: models achieve high performance but rely on spurious correlations rather than diagnostic criteria, while models that focus on micromorphologically relevant

Open Peer Review

Approval Status ✓ ✓ ? ✓

	1	2	3	4
version 2 (revision) 04 Aug 2026	✓ view	✓ view		✓ view
version 1 02 May 2026	? view	? view	? view	

1. **Mathias Bellat**, University of Tübingen, Tübingen, Germany
2. **Jerome G. Gacu**, Kyoto University, Kyoto, Japan
3. **Alexandra Karamitrou**, University of Southampton, Southampton, UK
4. **Anno Hein**, Institute of Nanoscience and Nanotechnology, National Centre of Scientific Research "Demokritos", Aghia Paraskevi, Greece

Any reports and responses or comments on the article can be found at the end of the article.

features show lower overall performance. Expert agreement on the same task is low, with uncertainty concentrated in feature detection rather than classification. Crucially, model and expert errors are largely independent, and each captures different aspects of frost feature recognition, establishing a basis for complementarity.

Conclusions

These findings demonstrate that effective computational integration in micromorphology requires not only accurate classification but interpretability validation ensuring that model's reason from the same diagnostic criteria as experts. We propose a human-in-the-loop approach where models provide consistent first screening, while experts offer contextual interpretation and diagnostic validation. Additionally, we present an interactive open-access tool that implements this pipeline to facilitate adoption and repeatability.

Plain Language summary

Traces of frost action in sediments, such as distinct arrangements of soil particles and pore spaces formed by repetitive cycles of freezing and thawing, are important indicators of ancient glacial environments and how humans adapted their behavior to past cold climates. Frost features are usually identified by experts examining thin sections of soil and sediment under the microscope, a process that takes time and relies heavily on individual experience. In this study, we explore whether artificial intelligence can assist experts in recognizing frost-related features in archaeological sediments. We trained models to classify images showing different types of frost structures and compared their results with those of experienced micromorphologists who evaluated the same images. The models performed as well as human experts and sometimes recognized features that experts found difficult to classify. However, experts still provided essential insights for interpreting the meaning and context of these patterns. Together, this collaboration between human specialists and artificial intelligence offers a new way to make sediment analysis faster, more consistent, and more reproducible across archaeological research.

Keywords

Micromorphology; Frost action; Deep Learning; Convolutional Neural Networks; Expert-Model collaboration

HE

This article is included in the Horizon Europe gateway.

This article is included in the European Research Council (ERC) gateway.

Corresponding authors: Sofia Kouki (skouki@ualg.pt), Vera Aldeias (vlaldeias@ualg.pt)

Author roles: Kouki S: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Li L**: Conceptualization, Methodology, Supervision, Validation, Writing – Review & Editing; **Aldeias V**: Conceptualization, Data Curation, Funding Acquisition, Project Administration, Supervision, Validation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No [101041245](Into the Sedimentary Matrix: Mapping the Replacement of Neanderthals by early Modern Humans using micro-contextualized biomolecules [101041245]).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2026 Kouki S *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 work is properly cited.

How to cite this article: Kouki S, Li L and Aldeias V. **Automated detection of freeze-thaw signatures in archaeological sediments using deep learning [version 2; peer review: 3 approved, 1 approved with reservations]** Open Research Europe 2026, **6**:144 https://doi.org/10.12688/openreseurope.23697.2

First published: 02 May 2026, **6**:144 https://doi.org/10.12688/openreseurope.23697.1

REVISED **Amendments from Version 1**

The revised version incorporates changes in response to three peer reviewers. The main revisions are as follows:

Text and structure: We reconstructed the introduction to improve the logical flow between the description of frost feature identification challenges and the rationale for computational approaches by adding a paragraph that bridges the two sections. We revised the abstract to clarify the two-stage classification approach. We expanded the Machine Learning workflow section with additional rationale for key method choices. The Expert-CNN Models Alignment results section was streamlined, and a figure (Figure 8) and a table (Table 2) were added to better support the section. We expanded the limitations and future work section to include explicit discussion of expert survey constraints, independent dataset validation, multi-scale architecture directions, and dataset splitting strategies.

Figures and Tables: Figure 1 (Figure 3 in the new version) was redesigned with a shared x-axis, increased size, and 95% confidence intervals replacing the standard deviation error bars. We added a new Figure 1 showing the geographical distribution of sites with representative photomicrographs. The workflow figure (previously in Supplementary Information 4) was incorporated into the main text as Figure 2. The previous Figure 2 (Figure 4 in the new version) was updated with the misclassification annotations. The previous Figure 3 (in the new version Figure 5) was updated with a single shared color bar, annotations of the diagnostic features. We moved figures (Figure 8) and tables (Table 2) from the Supplementary Information to better support the result section and improve the readability to readers non-familiar with the topic.

Supplementary Information: The glossary table (Table S4.1.1) was expanded to include formal equations for all performance metrics and statistical tests. The dataset labeling procedure was expanded and described in Section S4.2.

Any further responses from the reviewers can be found at the end of the article

Introduction

Much of our understanding of the past derives from the study of sediments and soils. The sedimentary record preserves a rich archive of information and, as such, constitutes a primary focus of research across numerous disciplines, including archaeology, soil sciences, and geology. Soil micromorphology, the microscopic study of undisturbed soil and sediment in thin sections, has become a cornerstone analytical method in these disciplines. By examining sediment composition, fabrics, pedogenic features and their subsequent micro-scale spatial relationships, micromorphologists reconstruct depositional environments, and site formation processes, revealing both natural environments and anthropogenic activities that shaped the geological and archaeological record (Courty, 1989; Goldberg et al., 2022; Goldberg and Aldeias, 2018; Karkanas and Goldberg, 2018; Nicosia and Stoops, 2017).

Despite its central analytical power, micromorphology faces three persistent methodological constraints. First, analysis is labor-intensive and time-consuming, often requiring months of specialist effort for comprehensive site studies. Second, interpretations depend heavily on expert experience and knowledge for recognizing diagnostic features through years of training; yet these features are often subtle, variably expressed, or context dependent. Third, this interpretive complexity creates persistent inter-observer variability, reducing reproducibility, and limiting our ability to build large, comparable datasets across sites and regions (Shahack-Gross, 2016). These challenges hinder cumulative comparisons, as quantitative syntheses are essential for reconstructing past environments and behaviors at regional scales.

One prominent example of these challenges is frost action, where repeated freeze-thaw cycles create diagnostic microscopic signatures, including granular, blocky, and platy-lenticular microstructures, capped grains, and characteristic porosity patterns (Van Vliet-Lanoë, 1998, 1985; Vliet-Lanoë, 2010). From an archaeological perspective, recognizing these frost signatures is critical for multiple reasons. First, they enable reconstruction of past climatic conditions and understanding of hominin adaptations to glacial and periglacial settings (Courty and Vallverdu, 2001; Davies et al., 2003; Dibble et al., 2017; Holcomb et al., 2023; Kehl et al., 2025; Murphree et al., 2025; Straus, 1991; Vachula et al., 2019; Van Vliet-Lanoë, 1998). Second, frost action affects the preservation of archaeological materials (Deeben et al., 2010; Texier et al., 1998; Todisco and Monchot, 2008) and disturbs the spatial integrity of previously deposited archaeological assemblages (Bertran et al., 2015, 2010; Masson, 2010).

Yet despite extensive characterization in archaeological contexts, frost features vary along a continuum from weakly developed to well-defined forms, their expression is dependent on textural features (e.g., grain-size), and they can be obliterated by or overlap with other pedogenic processes (Bertran et al., 2017, 2015; Phillips et al., 2013; Van Vliet-Lanoë, 1998, 1985; Van Vliet-Lanoë et al., 2004). These characteristics (variability along a continuum, textural dependency, and overlap with other processes) create the conditions where expert judgment becomes inconsistent and reproducibility suffers.

These challenges are well suited to image classification. Frost feature identification is a visual pattern recognition task operating on spatially defined features (aggregate geometry, porosity, and microstructure arrangement) across large sets of images. The features themselves, while subtle and variably expressed, are diagnostically grounded in established micromorphological criteria, providing a structured classification framework. At the same time, inter-observer variability documented above indicates that consistent, automated screening could meaningfully complement expert interpretation by providing a reproducible first-pass assessment across datasets that are too large for manual analysis. Indeed, computational approaches have been adopted in related fields to address similar interpretive challenges arising from subtle features, contextual variability, and observer subjectivity. In geological petrography, where similar interpretive challenges exist, Convolutional Neural Networks (CNNs) approaches have proven effective for diverse classification tasks, such as rock and mineral classification (Basso et al., 2025; Dalhat and Osman, 2025; Faria et al., 2025; Iyas et al., 2020; Maitre et al., 2019; Polat et al., 2021; Su et al., 2020), microfacies classification (Pires de Lima et al., 2020), fossil classification (Liu et al., 2023; I. C. Romero et al., 2020a), porosity and grain size analysis (Rubo et al., 2019; Tang et al., 2020).

However, CNNs are often treated as "black boxes", limiting their adoption in domains where diagnostic criteria are well established and where high accuracy does not necessarily indicate that models rely on meaningful features (Ebert-Uphoff and Hilburn, 2020). To address this, a range of interpretability methods have been developed to visualize the spatially discriminative regions that influence model predictions, including Class Activation Maps (CAMs; Barredo Arrieta et al., 2020; McGovern et al., 2019; Şahin et al., 2025) and perturbation-based approaches such as Shapley Additive Explanations (SHAP; Lundberg and Lee, 2017; Stadlhofer and Mezhuyev, 2023). These methods improve the justifiability and trustworthiness of computational approaches, making model decisions more transparent and interpretable (Bergen et al., 2019; Jiang et al., 2024; Reichstein et al., 2019). Several studies in geological petrography have incorporated interpretability analysis on their classification scheme, demonstrating both the utility and challenges of interpretability analysis (Bai et al., 2025; Seo et al., 2022; Zheng et al., 2024).

Despite these advances in geological petrography and the shared interpretative challenges, the application of machine learning to archaeological micromorphology remains remarkably limited. To date, Arnay et al. (2026, 2021) have demonstrated the viability of deep learning for archaeological thin section analysis, first for pore type classification and more recently for segmentation of bones, charcoals, and lithic materials and estimate their subsequent abundance in thin sections. Taking a distinct methodological approach, Zickel et al. (2024) developed a GIS-based toolbox that applies semi-supervised random forest classification, rather than deep learning, to multi-modal imagery in a multi-band raster for pixel-level classification and spatial quantification of sediment components. To our knowledge, no studies have yet addressed frost feature identification in archaeological sediments, compared models and expert performance on micromorphological classification tasks or used interpretability analysis to evaluate whether models learn micromorphologically meaningful features.

Here, we address this gap by evaluating machine learning and expert performance in parallel to the classification of frost-related features in archaeological thin sections. First, we develop and evaluate CNN-based models that identify and classify frost features in photomicrographs. Second, we apply interpretability analysis to visualize model attention patterns and assess whether models learn micromorphologically meaningful diagnostic criteria. Third, we quantify inter-expert variability in frost classification, establishing a baseline for practicing micromorphologists' performance. Fourth, we directly compare model and expert classifications to evaluate whether computational approaches provide complementary capabilities. Together, these analyses provide a framework for assessing how automated approaches can meaningfully support identification and classification of frost features in archaeological micromorphology.

Materials and methods

Site selection and data preparation

We selected thin section samples from Plio-Pleistocene archaeological sites spanning a broad range of geographical and environmental contexts, specifically targeting deposits containing frost-affected and non-frost-affected sediments. European sites included Quinçay and La Ferrassie (France), Ortvale Klde (Republic of Georgia), El Cierro, La Garma (Spain), Lapa do Picareiro (Portugal), and Bacho Kiro (Bulgaria). Sites beyond Europe included Boker Tachtit (Israel), Ali Toyta (Ethiopia), and Contrebandiers Cave (Morocco). We chose these sites because they represent different periglacial and non-periglacial depositional environments with the hope to build a model that can (later) be generalized across diverse archaeological settings (Figure 1).

We analyzed 154 thin sections, yielding 733 photomicrographs captured under plane-polarized light (PPL) using a Nikon SZM25 stereomicroscope equipped with a DS-Fi3 digital camera at 5x magnification and 1x zoom. At lower magnifications (e.g., scanned thin section images) the broader depositional context is visible, but individual diagnostic features,

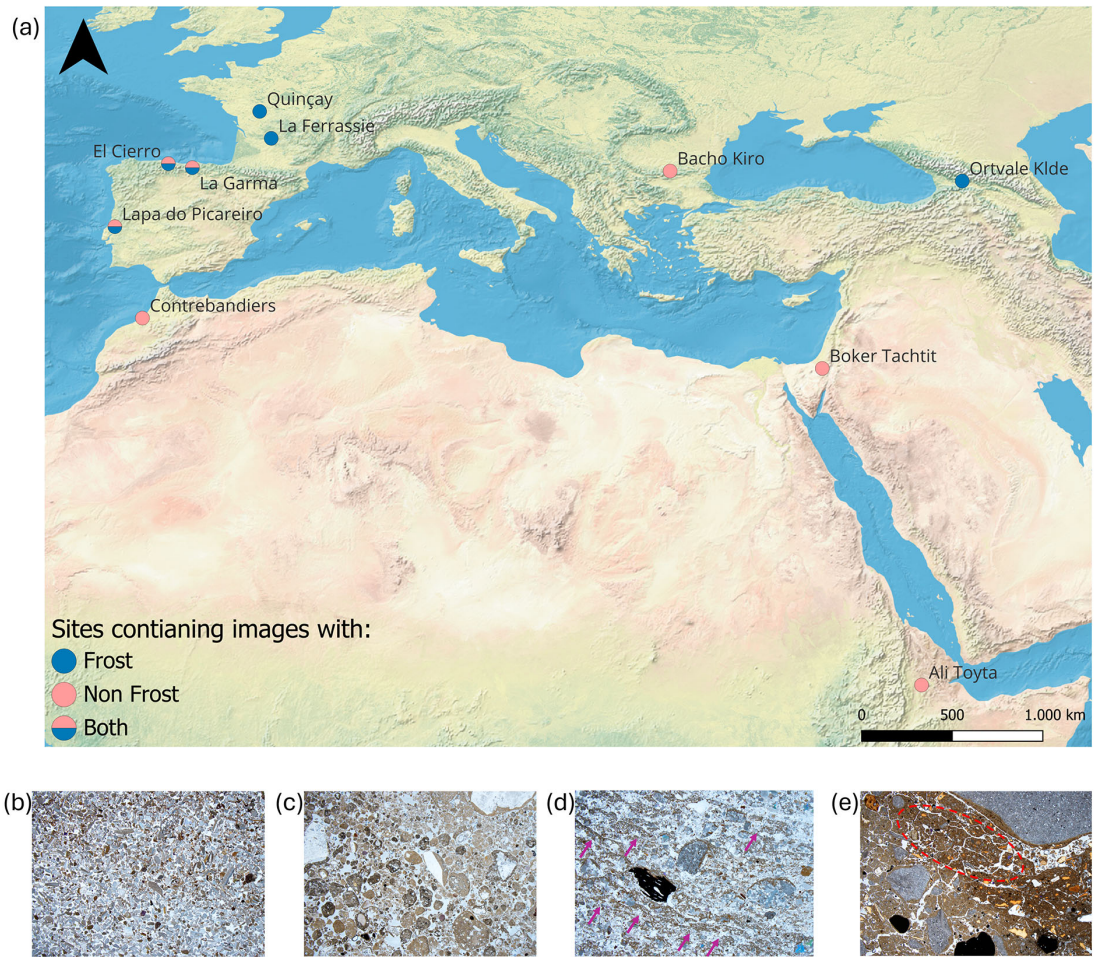

Figure 1. Geographic distribution of the sites and representative sample photomicrographs. (a) Map showing the sites used for frost feature identification and classification. Blue cycles indicate sites where images contain frost features (frost, and subsequent microstructure types: granular, blocky, platy-lenticular). Pink cycles indicate sites where images contain no frost features. Half-pink, half-blue cycles indicate mixed sites containing layers with either frost and non-frost features. Digital Elevation Model (DEM) data sourced from Natural Earth (public domain; https://www.naturalearthdata.com/). Map created using QGIS 3.40.6. (b-d) Thin-section photomicrographs. (b) Photomicrograph with non-frost features from Contrebandiers site showing geogenic bioclastic sands (Aldeias et al., 2014). (c) Photomicrograph from Quinçay showing frost features with granular microstructure. (d) Photomicrograph from La Ferrassie showing frost features with platy-lenticular microstructures (banded fabric; purple arrows) (Aldeias et al., 2023). (e) Photomicrograph from Ortvale Klde showing frost features with blocky microstructures. The dashed red ellipse shows sub-angular blocky microstructure. An interactive map showing representative photomicrographs from all sites is provided in data repository (10.17605/OSF.IO/TYDSM).

such as aggregate boundaries and shape, porosity distribution, lack sufficient resolution for the classification. At higher magnifications, fine details are resolved, but the spatial relationships between features that define the microstructure (granular, blocky, platy-lenticular) types fall outside the field of view. In our case, a 5x magnification provides a balance between the feature resolution and spatial context, which is necessary for the classification. We also used PPL entirely to create a controlled, single-modality imaging procedure, which is particularly useful since diagnostic criteria are directly observed under PPL. Also, merging PPL and XPL computationally would necessitate either multi-model fusion structures lacking proven methods in micromorphological imaging or the possibility of adding methodological artifacts through correlated duplicate observations in the same region.

All the images we acquired are in RGB format, with a of 2880 x 2048-pixel and saved as uncompressed .tif files. These raw images formed the input dataset for subsequent data preparation and model training described in Section 2.2.

We cropped each photomicrograph into six equal parts to capture local textural variability, and we excluded areas with no relevant sedimentary information (e.g., slide edges, large rock fragments). This allowed us to increase dataset size and ensure models learn from localized diagnostic features suitable for classification (Lakshmanan et al., 2021). After cropping, we resized the images to 512 x 512 (width x height) pixels, ensuring that all images are the same size before model implementation.

We established reference labels based on two annotation levels: (1) binary classification (Frost vs. non-frost), and (2) multiclass classification of Frost images into Blocky, Granular, or Platy-lenticular microstructure types following the standardized micromorphological nomenclature (Stoops, 2021), though simplified here to enhance the practical applicability.

Our dataset consist of 921 images in total and randomly split into training (70%), validation (15%) and test (15%) sets (Supplementary Information 4, Table S4.2.1). We augmented only the training set with random rotations, flips, and height/width shifts. This allows us to increase the training dataset images to avoid overfitting. Since some image classes contain more images than others, and prevent overconfident predictions on minority classes, we addressed this class imbalance by computing class weights and applying label smoothing (Müller et al., 2019; Spelmen and Porkodi, 2018).

Machine Learning workflow

We applied five CNN architectures for the two-stage classification framework: MobileNetV2, EfficientNetV2B0, ResNet50, a custom CNN, and a custom Bayesian Neural Network (BNN). On one hand, we selected three widely adopted CNN initialized with ImageNet weights models (MobileNetV2, EfficientNetV2B0, ResNet50) to leverage transfer learning, which provides low-level feature extraction (such as edges and textures) on small-to-medium datasets without requiring massive training volumes dataset (Lakshmanan et al., 2021). Specifically, MobileNetV2 emphasizes computational speed and has a simple feature extraction design by using inverted residuals (Sandler et al., 2019). ResNet50 uses a more complex layered structure with shortcut mappings to capture detailed patterns and prevent the vanishing gradient problem (He et al., 2015), while EfficientNetV2B0 provides a balanced approach by combining both computational efficiency and complex detection capability via optimized compound scaling (Tan and Le, 2021).

On the other hand, Custom CNN and BNN models were designed from scratch specifically for this task to force the networks attention onto domain-specific matrix textures rather complex, object-centric attention inherited from ImageNet pretraining. We selected BNN specifically because it allows for uncertainty estimation (confidence intervals) in predictions and to do so we used Monte Carlo (MC) dropout, executing 20 stochastic forward passes per image to quantify the epistemic uncertainty (model parameter variance) of our classifications (Jospin et al., 2022; Wang and Yeung, 2021).

To evaluate all models' generalization, we applied 5-fold stratified cross-validation (CV) with each fold presenting class proportions and ensuring that every image is used for validation once in one of the five subsets. We chose stratified over random cross-validation to maintain consistency across class proportions across all folds, reducing the variance in performance estimates, especially for minority classes (Ghojogh and Crowley, 2023; Gorriz et al., 2024). After completing CV, a final model for each one of the CNN-based models was trained on the entire augmented training set and evaluated on an independent held-out test set, providing an unbiased estimate of model performance on unseen data (Ghojogh and Crowley, 2023).

For all models, we used ReLU as the activation function to introduce non-linearity while avoiding saturation, followed by global average pooling and fully connected dense layers to reduce spatial dimensions and avoid overfitting, and fully connected dense layers to compute final class probabilities. To prevent overfitting, we applied high dropout rates (0.3–0.5) to randomly deactivate neurons during training steps, l2 regularization to penalize large weight values (Gavrilov et al., 2018), and adaptive learning rate scheduling to systematically reduce the step size as training converged on local minima (Santos and Papa, 2022; Ying, 2019).

We evaluated all models' performance using accuracy, precision, recall, F1-scores, and confusion matrices, as well as visualizing the misclassified images per model (Raschka, 2020; Tohka and van Gils, 2021). To clearly assess which model performed better and how large the performance differences were across both classification tasks, we compared metrics across all folds and the test set using the Friedman test with Nemenyi post-hoc tests to evaluate significance between models (Demšar, 2006a; Rainio et al., 2024).

To understand which parts of an image influences the model's decision-making process, we applied Score Class Activation Map (Score-CAM), a visualization technique that generates maps which highlight the image regions that mostly affect a model decision that an image belongs to one class or another (Minh, 2023; Wang et al., 2020). To see if

these models can be used for archaeological deployment and if the models learn meaningful patterns, we also qualitatively compared if these activation maps align with the established micromorphological criteria, such as frost-related porosity, aggregate shape and arrangement, or microstructure geometry (Stoops, 2021; Van Vliet-Lanoë, 1985; Vliet-Lanoë, 2010). In addition to qualitative inspection, we quantified Score-CAM attention concentration by summarizing what proportion of pixels is needed to reach fixed activation mass thresholds per model and per image. Full implementation details are provided in Supplementary Information 4 (Section S4.2.2).

We determined optimal hyperparameter values for learning rates, dropout rates, regularization coefficients, and models' architecture-specific parameters (e.g., number of dense layers) through systematic iterative experimentations. We tested variations of these hyperparameter values, evaluated performance on the validation set, and refined settings based on results. This process balanced multiple objectives: maximized accuracy while preventing overfitting and ensuring stable convergence. The final hyperparameter configurations for all models are reported in Supplementary Information 4. Table S4.2.2.

Since the random splitting operates at crop level, some crops form the same photomicrograph may appear in different sets. However, in our case, the photomicrograph-level split was not feasible because it would have drastically reduced sample size given the already limited dataset. To assess whether the crop-level splitting substantially inflates performance estimates, we performed a small robustness test on full photomicrographs (n = 21; 11 frost, 10 non-frost), applying all five models and computing the same metrics used for the cropped images, together with Score-CAM maps to assess the whole-image behavior (See Supplementary Information 4, Section S4.3 for details).

Expert survey design

Model performance metrics alone cannot assess whether experts would agree with model classifications or whether the reference labels themselves are reliable. Therefore, we conducted an anonymous online survey among practicing micromorphologists (Microsoft Forms v25.7), distributed via direct invitations to micromorphologists and remained open from May 10, 2025, to October 31, 2025.

The survey serves three purposes: (1) quantify inter-expert agreement to identify potential label noise in the training data, (2) establish whether certain images are inherently ambiguous even to experts, and (3) provide a baseline for further comparison between CNN performance to human expert performance.

We selected a subset of 23 images that are representative of our dataset in a two-stage process. First, we identified potentially ambiguous images based on frequent misclassification by preliminary CNN models, (misclassified by ≥3 of the models), and cases where reference labels require extended expert discussion without reaching initial consensus. Second, we refined this subset through expert review to ensure representation of specific ambiguity types, such as poorly defined frost features, frost features not dominating the field of view, and non-frost images displaying potentially confounding characteristics.

In the first task (first screening assessment, n = 9), participants evaluated whether the sediment features within the resized image are sufficient to attempt classification. We selected these 9 images specifically because preliminary CNN models frequently misclassified them and because reference label assignment required extended discussion, suggesting that diagnostic features were difficult to identify at the cropped and resized scale.

In the second task (binary and multiclass classification, n = 14), participants classified images in a two-stage process mirroring the CNN framework: first assessing whether images contain any frost features, and if they do, then the specific type of microstructure. This conditional design prevents forced classification and allows direct comparison with model outputs.

We assessed inter-rater agreement using Fleiss' kappa, individual expert-reference labels alignment using Cohen's kappa (Landis and Koch, 1977), raw percentage agreement, and relationships between response variability and agreement using Spearman's ρ (Mukaka, 2012). We also quantified the proportion of "Unsure" responses per participant and per image. Full statistical procedures are detailed in Supplementary Information 4, Section S4.4.

Experts and CNN models alignment

To evaluate whether CNN-based models align with expert micromorphological practice, we compared model predictions and expert classifications across the 14 images from the survey's second tasks, the images where both experts and models completed full two-stage classification. All comparisons used the reference labels as reference.

Evaluating experts and CNN models' alignment addresses three questions: (1) do CNN models agree with expert classifications, (2) do they make similar errors on the same images, and (3) does model uncertainty correspond with expert-perceived classification difficulty.

To evaluate how models behave on images where experts judged diagnostic features as insufficient for classification, we used the nine images from the survey's first-screening task. Since these nine images presented challenges in feature visibility at the cropped and resized scale, they were exclusively for assessing whether models can distinguish between images with sufficient and insufficient diagnostic information. The remaining 14 images, where features were considered sufficiently visible for classification, were used for the two-stage classification tasks (binary and multiclass) and form the basis of the expert-model comparison described below. For the nine first-screening images, we categorized each by majority expert vote as "Visible", "Not visible", or "Unsure". For each model, we generated frost/non-frost prediction on all nine images and summarized (i) class proportions on "Not visible" images, (ii) mean confidence scores (Equation 1, Supplementary Information 4, Section 4.5) for "Visible" vs. "Not visible' subsets, and (iii) For the BNN, Monte Carlo dropout uncertainty (standard deviation over 20 passes).

For the classification task, we assessed the overall degree of agreement using Fleiss' kappa first for the CNN models only ($n = 5$) and then all participants combined ($n = 24$; 5 models and 19 experts). We also computed Cohen's kappa for all 95 expert-model pairs and classification accuracy against reference labels for each participant. To test whether experts agreed more with one another than with models, we compared Cohen's kappa distributions using Mann-Whitney U test (Conover, 1999), with the same comparison applied to accuracy values. We quantified the error overlap between experts and models using Spearman's ρ between misclassification patterns. We also used Spearman's ρ to examine relationships between agreement proportions, model confidence scores, and BNN uncertainty metrics.

We categorized images into difficulty categories based on classification performance: (1) "Hard for CNN" correctly classified by $\geq$70% of experts but misclassified by $\geq$3 models; (2) "Hard for Experts" correctly classified by $\geq$3 models but correctly classified by <70% of experts; (3) "Hard for Both" misclassified by $\geq$3 models and correctly classified by <70% of experts. These thresholds were selected to identify images with clear difficulty patterns while maintaining sufficient sample sizes per category. Complete statistical pipelines and details are provided in Supplementary Information 4.

Workflow and interactive classification tool

The entire workflow, including image pre-processing, classification, evaluation, and interpretability were implemented in Python 3.12 with Tensorflow 2.20.00 (Abadi et al., 2016) and Keras 3.12.2 (Chollet, 2015) libraries within Google Colaboratory (Bisong, 2019) using an A100 GPU. For the three off-the-shelf models, ImageNet pretrained weights were loaded directly through Keras API (MobileNetV2: Sandler et al., 2019; EfficientNetV2B0: Tan and Le, 2021; ResNet50: He et al., 2015).

Statistical analyses of the survey data and model comparisons were performed using Scikit-learn (Pedregosa et al., 2011) and Statsmodels (Seabold and Perktold, 2010). Our workflow is presented in Figure 2 and a detailed glossary of the machine learning and statistical terms used in this study is included in Supplementary Information 4, Table S4.1.1.

To support practical adoption and reproducibility, we also developed an interactive web application using Gradio (Abid et al., 2019) and Hugging Faces that enables users to upload their thin section photomicrographs and obtain frost classification predictions with confidence scores, uncertainty and interpretability, implementing the same two-stage classification framework (binary followed by multiclass). The tool is publicly available at online (https://huggingface.co/spaces/skkou/frost_classifier) and detailed guidelines on how to use it are provided in Supplementary material 4, Section S4.8.

Results

CNN models achieve high performance but rely on spurious correlations

Overall, all five models (MobileNetV2, EfficientNetV2B0, ResNet50, Custom CNN, and BNN) achieve high classification performance both binary (frost vs. non-frost) and multiclass (blocky, granular, and platy-lenticular) tasks in both 5-fold cross-validation and the test set, with accuracy ranging from 94-99% in binary and 86-96% in multiclass. Models achieve better performance on binary tasks, while distinguishing among the three frost-related microstructures is more challenging (see Supplementary Information 1 and 4; Section S4.6–7).

Learning dynamics shows stable convergence for all models in the binary task, while there are more variable validation patterns with oscillations during the training procedure in the multiclass task (Supplementary Information 4, Figure S4.6.1-S4.6.3). Cross-validation results (precision, recall, and F1-score) are summarized in Figure 3. In binary

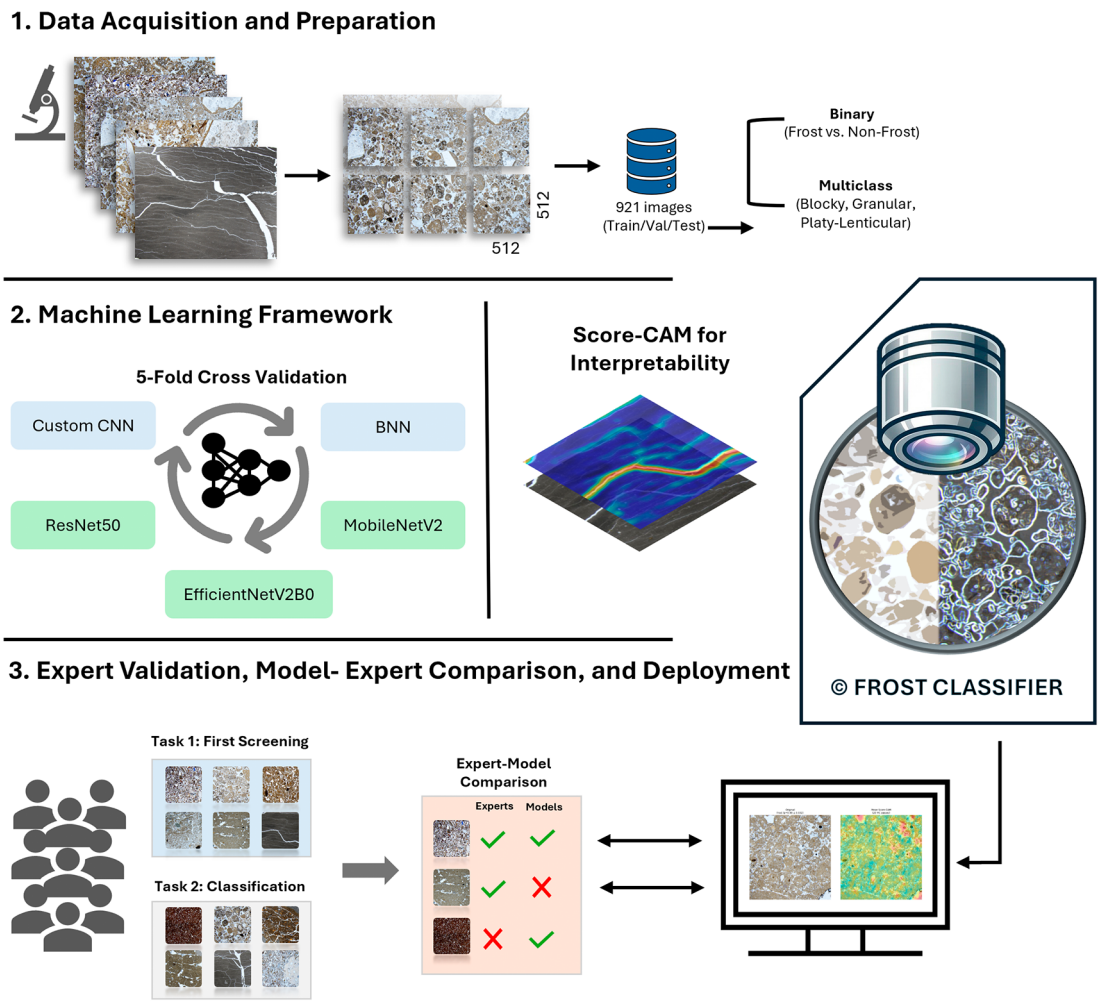

Figure 2. Frost classification workflow. End-to-end methodology, starting from the data acquisition and acquisition, model training and validation, expert survey and comparison with model outputs, and final deployment of the classifier as interactive tool.

classification, MobileNetV2 and EfficientNetV2B0 achieve the highest mean accuracy (98–99%), with all other models above 94%. In multiclass classification, performance is slightly lower. MobileNetV2 performs best (~ 96%), followed by EfficientNetV2B0 and ResNet50 (~ 94–95%). Images with granular features are classified consistently across all models (>93% precision and recall), while images with blocky and platy-lenticular features show greater variability. Evaluation of the test set confirms these cross-validation patterns, where MobileNetV2 and EfficientNetV2B0 achieve the highest performance on binary classification and show slightly lower performance in multiclass classification (Supplementary Information, Tables S4.6.1–2). Statistical comparisons using Friedman's test confirm significant differences in binary classification, with MobileNetV2 and EfficientNetV2B0 achieving the highest ranks. No significant differences are detected in the multiclass classification (Table 1A).

Misclassification patterns are consistent in both binary and multiclass classification (Figure 4). In the binary classification, non-frost images contain high spatial heterogeneity, while in the frost class and the subsequent microstructure classification, misclassifications are concentrated in images where the diagnostic features are either poorly developed, densely packed or occupy less than 30% of the image area.

To assess whether models rely on micromorphologically diagnostic features (such as aggregate geometry, frost-related porosity, and sediment arrangement), we applied Score-CAM interpretability analysis. For both classification tasks, MobileNetV2, EfficientNetV2B0, and ResNet50 models focus on scattered or fixed activation patterns that do not align with diagnostic features used by micromorphologists (Figure 5). ResNet50 consistently highlighted a central rectangular

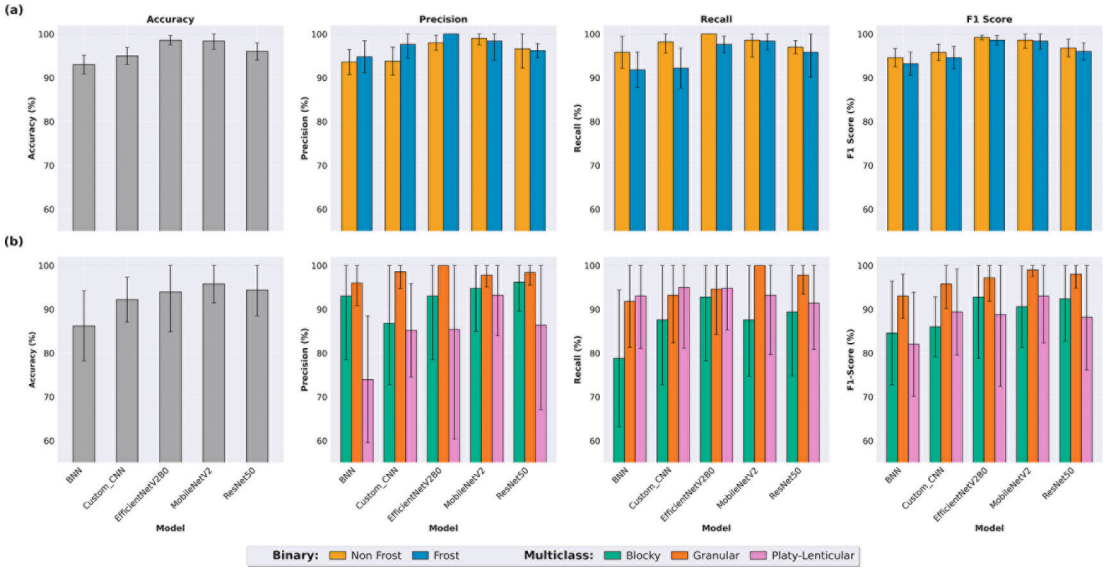

Figure 3. Comparative performance of five CNN architectures (BNN, Custom CNN, EfficientNetV2b0, MobileNetV2, and ResNet50) based on mean Accuracy, and per class Precision, Recall, and F1-Score and the standard deviation over 5-Fold cross-validation (CV). (a) Binary classification (Frost vs. non-frost) performance comparison. (b) Multiclass classification performance across Blocky, Granular, and Platy-lenticular microstructure classes. Error bars represent the 95% confidence intervals.

Table 1. Statistical comparison of CNN model performance and expert-model agreement across 14 classification images. (A) Model ranking based on Friedman's test with significant pairwise differences identified by Nemenyi post-hoc test. (B) Pairwise agreement comparison between expert-model pari and expert-expert pairs, assessed using Mann-Whitney U test. (C) Group-level agreement using Fleiss' kappa in three configurations (inter-expert, inter-model, and models-experts).

A. Model ranking and significant pairwise differences		
Model	**Rank**	**Significant differences**
MobileNetV2	1	>BNN (p = 0.039); >ResNet50 (p = 0.001)
EfficientNetV2B0	2	>ResNet50 (p = 0.005)
Custom CNN	3	-
BNN	4	-
ResNet50	5	-

B. Expert-model vs expert-expert pairwise agreement				
Comparison	**Cohen's kappa (SD)**	**Accuracy (SD)**	**N pairs**	**Statistical test**
Expert-Model	0.387 (0.343)	70.4 (21.1)	95	U = 7397, p = 0.224
Expert-Expert	0.447 (0.361)	74.1 (21.2)	171	Reference
Δ (difference)	−0.060	−3.7	-	Not significant

C. Group agreement configurations				
Configuration	**Fleiss' kappa**	**Mean agreement (%)**	**N raters**	**Interpretation**
Inter-Expert	0.202	60.9	19	Slight
Inter-Model	0.676	91.4	5	Substantial
Models-Experts	0.212	64.6	24	Fair

Note: Friedman x2 = 33.4, p = < 0.0001. Nemenyi post-hoc test, a = 0.05.

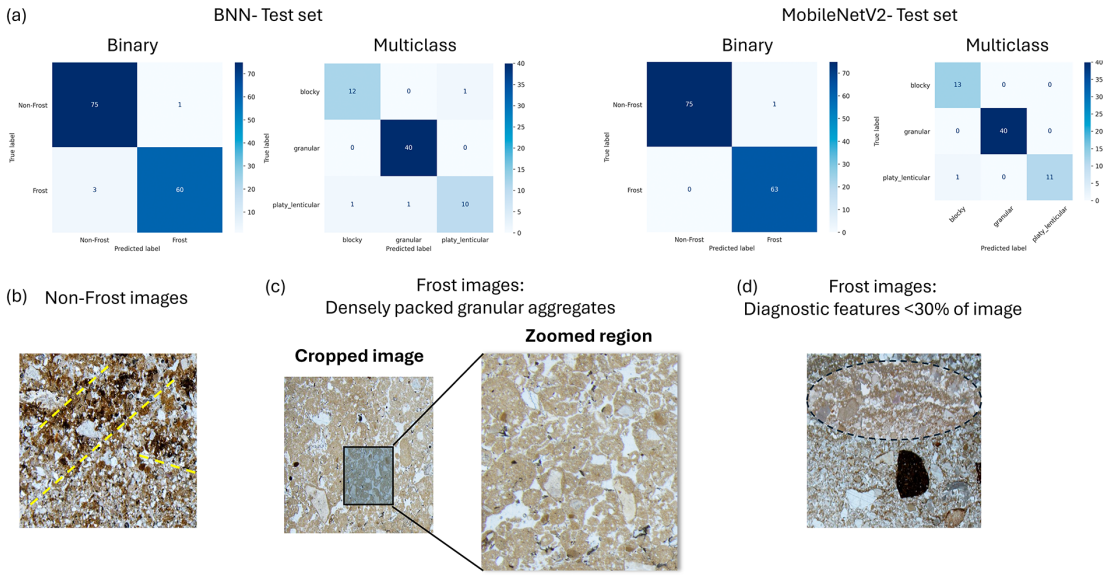

Figure 4. Systematic misclassification patterns across CNN models and representative examples of challenging images. (a) Confusion matrices on the test set for the best-performing off-the-shelf model (MobileNetV2) and a custom model (BNN) for binary (Frost vs. Non-frost) and multiclass (granular, blocky, platy-lenticular) classification tasks. (b) Representative Non-Frost images exhibiting high spatial heterogeneity. The complex mixed matrix with no dominant textural patterns mimics frost patterns (areas between the dashed yellow lines). (c) Representative frost images with densely packed granular aggregates where frost-related pore space (white) is obscured by tightly packed aggregates, as shown in the zoomed inset. (d) Representative frost images where diagnostic features occupy less than 30% of the image area. The dashed orange ellipse indicates the spatial extent of the diagnostic region within the cropped photomicrograph.

region regardless of the image content. In contrast, Custom CNN and BNN show more interpretable activation patterns (>0.6). These models show higher attention to aggregate geometry and characteristic porosity, especially in granular and platy-lenticular classes. However, both models occasionally attended to irrelevant elements within the image, such as bone fragments. The cumulative distribution of Score-CAM activation values across pixels shows that off-the-shelf models have highly concentrated attention with the top ~10–25% of pixels accounting for roughly half of the activation mass. Custom CNN and BNN require ~40–45% of pixels to reach the same threshold (Supplementary Information 4, Figure S4.6.6).

Quantitative Score-CAM analysis shows differences in attention concentration across architectures. Off-the-shelf models (MobileNetV2, EfficientNetV2B0, ResNet50) require only ~10–20% of pixels to reach 50% of activation mass, while Custom CNN and BNN needed approximately 40–45% of pixels showing more diffuse patterns (Supplementary Information 4, Figures S4.6.7–8 and Table S.4.6.1).

The complementary robustness test on the entire photomicrographs shows that off-the-shelf models, and particularly MobileNetV2, retained relatively balanced frost/non-frost performance. Whereas the custom CNN and BNN achieve lower performance metrics, but focused more consistently on micromorphological relevant features, with all models more prone to misclassify frost than non-frost images at whole scale (Supplementary Information 4, Figures S4.6.11–12 and Table S4.6.3).

Expert agreement on frost identification and classification is low

Of the 20 practicing micromorphologists who responded to the survey call, 19 completed it, with 4 shelf-reporting little to no experience, 10 having some experience, and 5 having significant experience with frost action. Due to small and uneven sample sizes, the statistical analysis performed has a more exploratory and descriptive intention rather than a definitive and inferential one. The survey's raw data and extended results are provided in Supplementary Information 2 and 4.

Overall, agreement among experts is low in both tasks. Fleiss's κ indicated only slight agreement ($\kappa \leq 0.20$; McHugh (2012)) (Table 1C). Figure 6a shows the broad spread of responses in both visibility and classification tasks relative to reference labels. The first-screening tasks produced roughly twice as many "Unsure" responses as the classification tasks, showing that experts find it harder to determine a feature that is present in the cropped and resized photomicrographs than to identify its class.

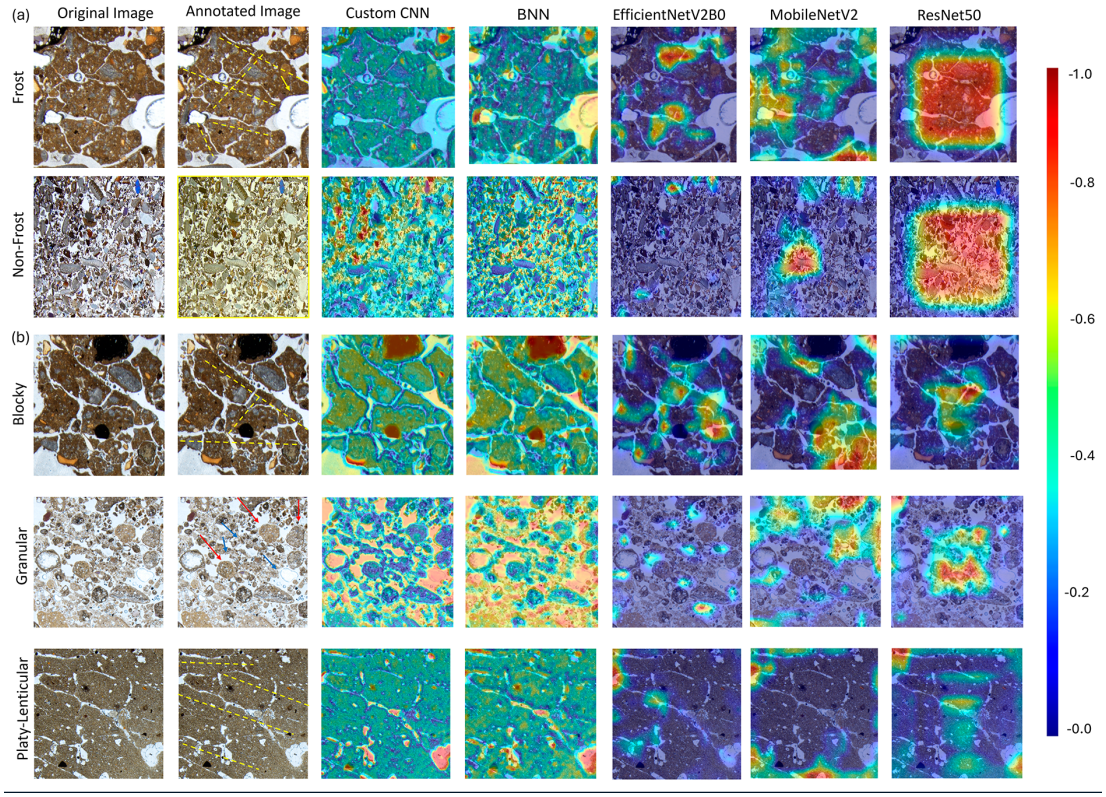

Figure 5. Score-CAM visualizations across five CNN architectures illustrating model attention patterns during (a) binary (Frost vs. non-frost) and (b) multiclass (Blocky vs. Granular vs. Platy-lenticular) classification. For each class, the same image is used across all five models to allow direct comparison of attention patterns. Original images are in PPL, where the porous space is white. Each heatmap highlights the image regions contributing most to model predictions. Differences in activation distribution reveal model-dependent focus and spurious correlations (for EfficientNetV2B0, MobileNetV2, and ResNet50), whereas other models (BNN and Custom CNN) emphasize more on diagnostically relevant areas (e.g., porosity) consistent with micromorphological criteria.

Individual performance varies widely. For the classification task, Cohen's κ relative to reference labels range from near chance to moderate agreement (Landis and Koch, 1977; McHugh, 2012), and percentage agreement per image ranges from 7 to 71%. In the first-screening tasks, percentage agreement ranges from 11 to 89%, but Cohen's κ could not be calculated because all reference labels are identical (assuming that all features are visible; Supplementary Information 4, Section S4.7). Frost-specific experience does not follow linear patterns. Experts with "some experience" show the highest median Cohen's κ (~ 0.45), while those with little to no experience (~ 0.20) and "significant experience" (~ 0.17) score lower with greater variability, producing an inverted U-shape pattern (Figure 6b).

Experts select "Unsure" for approximately 25–33% of images on average, with large individual variation (Supplementary Information 4, Table S4.7.4). In the first-screening task, uncertainty is negatively correlated with performance (Spearman $r = -0.800$, $p < 0.001$; Figure 6c), while no correlation is observed for the classification task (Spearman $r = -0.012$, $p = 0.961$; Figure 6d).

Models and experts show complementary strengths in frost classification

To compare models with expert performance, we evaluated how models' predictions aligned with reference labels and with the 19 experts who classified the same images in the survey's first screening and classification task. Extended results are provided in Supplementary Information 3 and 4.

On the nine first-screening images, models classified images confidently even when experts judged the images as insufficient for classification. Across the four images categorized as "Not visible" by majority vote, most models predicted an even split between frost and non-frost (2 each), with the exception of EfficientNetV2B0, which predicted frost for three of the four images. Mean confidence scores were higher for "Visible" images than "Not visible", and confidence remained high than dropping to chance level, with BNN showing the clearest differentiation (Table 2 and

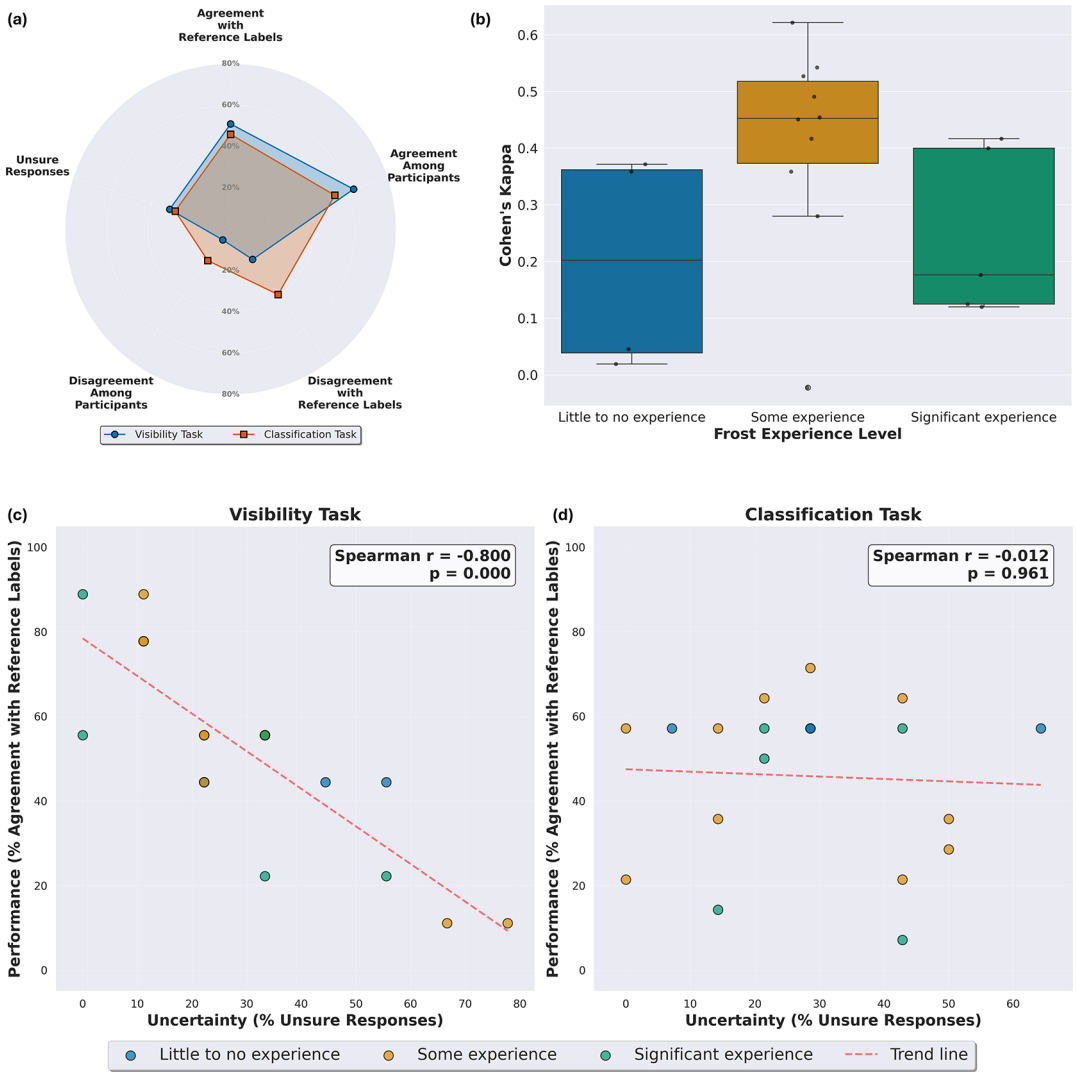

Figure 6. Summary of survey results evaluating expert reliability in frost identification and classification. (a) Comparison of response patterns between first screening (in the legend is indicated as visibility task) and classification tasks across five reliability metrics (0–80% scale): agreement with reference labels, inter-participant consensus, disagreement with reference labels, disagreement among participants, and unsure response frequency. (b) Distribution of inter-rater agreement (Cohen's Kappa) against the ground truth across frost experience levels during classification tasks, showing an inverted U pattern where participants with moderate frost experience have higher consistency among participants than participants with little to no experience and significant experience with frost-action. (c) Relationship between uncertainty (% of unsure responses) and performance (% of agreement with ground truth) for the first-screening task, colored by frost-experience level. The red dashed line shows downward trend, indicating strong negative correlation (Spearman r = −0.800, p < 0.001). (d) Relationship between uncertainty and performance for the classification task, colored by frost-experience level. Flat red dashed line indicates no correlation (Spearman r = −0.012, p = 0.961).

Table 2. Mean confidence scores and predicted mean probability for each model on the nine-first screening images, categorized by expert majority vote. Confidence is computed as the absolute rescaled distance from the decision boundary. P mean is the mean predicted probability of frost across images in each category.

Model	Not visible		Visible	
	Confidence	P mean	Confidence	P mean
Custom CNN	0.722	0.616	0.954	0.733
BNN	0.516	0.707	0.863	0.754
MobileNetV2	0.557	0.637	0.852	0.741
EfficientNetV2B0	0.860	0.776	0.984	0.751
ResNet50	0.870	0.548	0.937	0.744

Figure 8). In the classification task, model agreement with reference labels varies significantly (Friedman $x^2 = 33.4$, $p < 0.001$; Table 1A). MobileNetV2 achieved the highest alignment with reference lables (Cohen's $\kappa = 0.536$) and EfficientNetV2 performed well ($\kappa = 0.490$), with both significantly outperforming ResNet50.

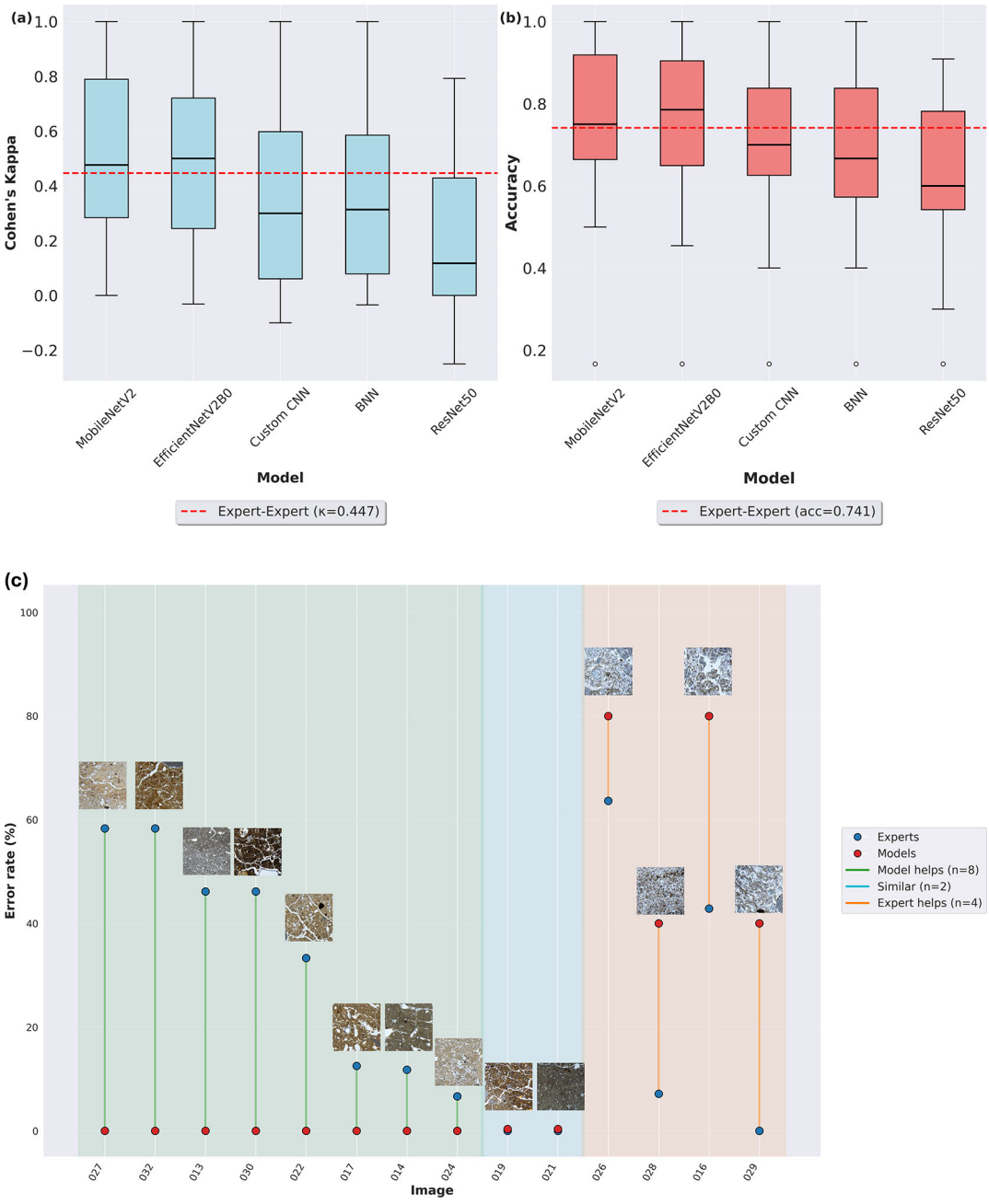

Figure 7. Expert-model agreement and image-level complementarity analysis. (a-b) Distribution of Cohen's kappa (a) and accuracy (b) for each CNN model against 19 experts. Red dashed lines indicate the expert-expert baseline ($\kappa = 0.447$, accuracy = 0.741). MobileNetV2 achieves the highest alignment with experts, approaching the expert-expert baseline. (c) Comparison of expert error rates (blue dot) and model error rates (red dot) for each of the 14 classification images, with thumbnail previews. Lines connect expert and model error rates per image, colored by complementarity patterns: green lines indicate images where models outperform experts (n = 8), cyan lines indicate similar performance (n = 2), and orange lines indicate images where experts outperform models (n = 4). Background shading corresponds to these categories.

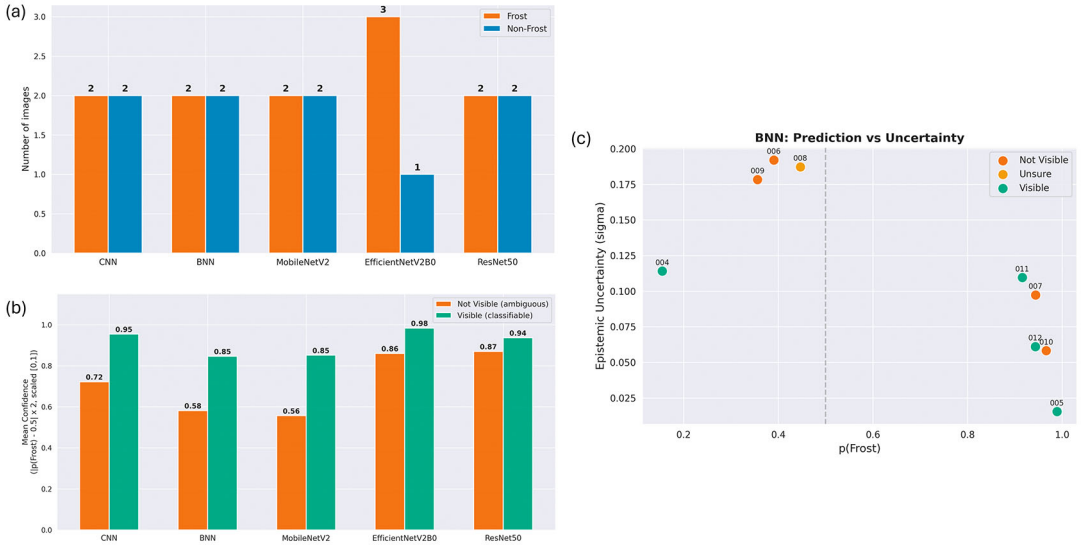

Figure 8. Model predictions, confidence, and epistemic uncertainty on the first-screening images. (a) Frost vs. non-frost predictions per model on the four images classified as "Not visible" by expert majority vote. (b) Mean confidence scores per model on "Not visible" and "Visible" image subset. Confidence scores are computed as rescaled distance from the decision boundary. (c) BNN predicted probability p(Frost) against epistemic uncertainty (σ, T = 20 Monte Carlo dropout passes) for all nine first-screening images, colored by expert visibility category. The dashed vertical line marks the 0.5 decision boundary between frost and non-frost classes.

Pairwise expert-model (95 pairs) and expert-expert (171 pairs) comparisons showed 70% and 74% accuracy respectively (Figure 7a-b; Table 1B). Differences are not statistically significant for either κ ($p = 0.224$) or accuracy ($p = 0.125$), showing that CNN models achieve expert-level inter-rater reliability.

Agreement patterns differ within groups (Table 1C). The five models show substantial consensus agreement (Fleiss' $\kappa = 0.676$, >90% mean agreement), exceeding inter-rater agreement (Fleiss' $\kappa = 0.202$, ~ 61% mean agreement; McHugh (2012)). When models are combined with experts (n = 24), Fleiss' κ (0.212, fair agreement; (Landis and Koch, 1977) approaches expert-only levels.

Image-level analysis reveals independent error patterns, where models and experts fail on different images, demonstrating complementary capabilities (Figure 5c). Eight images are difficult for experts, and models help in classification; four images need expert verification in the models' prediction, and in two models, experts' and models' classifications align. Finally, BNN uncertainty shows an inverse relationship with expert-perceived difficulty (Spearman $\rho = -0.497$, $p = 0.071$). Where experts expressed the highest uncertainty (~32-63% "Unsure" rate), BNN uncertainty remained low (Figure 8c).

Discussion

This study evaluates whether CNN-based classifications can identify frost features in archaeological thin sections with accuracy and reasoning comparable to practicing micromorphologists. We assessed model performance and interpretability, quantified expert variability, and directly compared models and expert classifications to determine whether computational and human approaches provide complementary capabilities. This framework enables us to evaluate not only classification accuracy but also decision mechanisms and the conditions under which CNNs can meaningfully support micromorphological practice.

We report four main findings: (1) CNN models achieve high classification performance, but best-performing off-the-shelf models rely on spurious correlations rather than diagnostic micromorphological features; (2) expert agreement on the same classification task is low, reflecting inherent interpretive challenges rather than methodological failure; (3) models and experts make independent errors on different images, indicating complementary strengths rather than redundant capabilities; and (4) model uncertainty does not align with expert-perceived difficulty, highlighting the need for expert oversight in hybrid workflows.

All models performed better on binary (frost vs. non-frost) than multiclass (blocky, granular, platy-lenticular) classification, and off-the-shelf models (MobileNetV2, EfficientNetV2B0, and ResNet50) performed better than those trained from scratch (Custom CNN, and BNN), a result that is consistent with the advantages of transfer learning for small-to-medium datasets reported in geological and petrographic imaging (Cai et al., 2024; Dawson et al., 2023; Pires de Lima et al., 2020). Among three widely adopted CNN architectures, lightweight models (MobileNetV2 and EfficientNetV2B0) generalized more effectively that heavier architectures (ResNet50), likely because their simpler designs are less prone to overfitting on limited training data (M. Romero et al., 2020b; Zhao et al., 2024). A similar pattern was also observed in carbonate rock classification where more complex CNN architectures performed better with bigger datasets (Dawson et al., 2023). The performance drop from binary to multiclass classification reflects a fundamental challenge, and several factors contribute to this performance decline. First, the classification threshold in the binary is easier to learn, and multiclass classification threshold boundaries might be vague or overlap with another class (Moral et al., 2022). Second, frost microstructures form a visual and functional continuum rather than discrete categories. Blocky, granular and platy-lenticular forms can be transitional, poorly developed or overlapping with other pedogenic processes through equifinality, for example granular microstructures from freeze-thaw cycles can resemble those produced from bioturbation (Van Vliet-Lanoë, 1985). This overlap between categories contrasts with machine learning techniques that demand discrete categorical boundaries, creating an epistemological mismatch between natural variability and the fixed category labels required for classification (Petrelli, 2024).

Furthermore, interpretability analysis showed that classification accuracy did not necessarily reflect the use of meaningful micromorphological features. Score-CAM maps showed that the highest-accuracy off-the-shelf models (MobileNetV2, EfficientNetV2B0, ResNet50) focused on scattered or central image regions unrelated to diagnostic criteria. This pattern is consistent with shortcut learning where deep networks exploit spurious regularities rather than focusing on the actual sediment features (Adebayo et al., 2020; Geirhos et al., 2020; Rosenfeld et al., 2018). In our case, the spatial attention patterns observed in off-the-shelf models likely reflect biases inherited from ImageNet pretraining, where images are predominantly object-centric and models focus on the part of the image with the most variability (Zhang et al., 2022). In ResNet50, the repeated activation of a central rectangular region across all photomicrographs exemplifies this center-of-the-photo bias. The additional quantitative Score-CAM analysis we conducted confirms that decisions in these off-the-shelf models are driven by compact regions of the image that only partially (or do not) overlap with micromorphological diagnostic sediment features, such as porosity and matrix arrangement, highlighting that high performance can coexist with non-diagnostic attention patterns.

In contrast, Custom CNN and BNN showed lower performance (varying from 86–95% in accuracy) but had clearer interpretability. Quantification of the pixel activation mass showed that the attention is more diffused, and Score-CAM overlays emphasized mostly on sediment matrix textures, frost-related porosity, and aggregate boundaries and shape, indicating that simpler architectures may force attention onto visual features that align with diagnostic criteria (Lipton, 2017; Rudin, 2019). This interpretability-performance trade-off suggests that simpler architectures focus on diagnostic features but lack representational capacity to integrate them across the frost pattern for robust generalizations (Raghu et al., 2019). At the same time, difference in feature extraction design across architectures (inverted residual blocks in MobileNetV2 (Sandler et al., 2019), compound-scaled blocks in EfficientNetV2B0 (Tan and Le, 2021), residual and squeeze and excitation blocks in ResNet50 (He et al., 2015), and the more linear custom CNN and BNN) and the specific last convolutional layer used for Score-CAM introduce distinct inductive biases in which regions are treated as informative.

In addition, the crop-level splitting may contribute to the optimistic performance estimates through spatial autocorrelation between crops from the same photomicrograph. However, the robustness test on the entire photomicrographs, which is independent splitting strategy, reproduces the same architecture-dependent pattern: off-the-shelf models retain better classification performance while custom models show better interpretability alignment. This suggests that the core findings are not primarily by crop level autocorrelation.

Therefore, high classification performance alone is insufficient for evaluating models in micromorphological contexts, where feature identification must be grounded in diagnostic criteria. Interpretability analyses are then essential not merely to explain model behavior, but to verify that classifications reflect the same process-based reasoning that experts apply (Ebert-Uphoff and Hilburn, 2020; Rudin, 2019). Similar interpretability paradoxes have been reported in geological petrography, where well-performing CNN's focus on visually irrelevant or non-diagnostic regions (Cai et al., 2024; Zheng et al., 2024). In other words, without explicit interpretability validation, deep learning models in micromorphology may be "right for the wrong reasons" and thus unreliable for process-based archaeological interpretations.

Interestingly, our survey results showed only a slight expert agreement both on ambiguous images and with reference labels. This reflects how micromorphological analysis typically unfolds: dynamically scanning thin sections while

switching between cross-polarized and plane-polarized light, changing magnifications, rotating stages to observe crystal orientations, checking the overall geological context. In the survey, we effectively removed these workflows by providing static images at fixed lighting and resolution without a more general geological provenience context. Furthermore, many of the photomicrographs provided to the experts had weakly developed frost features or features that mimic those related to frost action (e.g., developed pedogenic pedality). Experts therefore lacked crucial information for confident classification, explaining low agreement, high uncertainty rates, and disagreement with the reference labels.

These constraints intersect with deeper challenges. Frost features can overlap with other pedogenic (planar voids and peds in vertisols) or biogenic (coalescent granular structures derived from bioturbation) processes, producing microstructures that can mimic those typically derived from frost action, particularly at the photomicrograph scale (Stoops, 2021). This overlap is compounded by the continuum-to-category mismatch discussed above, which amplified the subjectivity in expert judgement (where to draw boundaries) and rigidity in model classification framework, which must assign a single category. Shahack-Gross (2016) showed similar low agreement among micromorphologists when looking at basic component and mineralogical identification, suggesting an interpretive variability in micromorphology analysis. Such subjectivity arises because micromorphological features vary in degree of development (from incipient to well-developed structures), diagnostic criteria involve development of quality judgments, and expert experience shapes recognition patterns.

This inherent subjectivity, however, also explains why model-experts complementarity is advantageous. The near-zero error correlation between models and experts indicates that machine versus human expert errors arise from distinct heuristics: where CNN models ask, “what does this look like?”; experts ask, “how did this form?”. These are different epistemic questions, meaning that CNNs are not simply faster versions of experts but different kinds of observers. Complementarity arises not from matching abilities but from non-overlapping strengths. Because models and experts succeed under different conditions, hybrid workflows are justified (Dvijotham et al., 2023). Once validated, CNNs can rapidly screen large image sets, provide consistency unaffected by fatigue, and flag uncertain cases. Experts can then adjudicate ambiguous cases, provide contextual interpretation, and integrate into sedimentary and paleoenvironmental narratives; neither replaces the other (Hemmer et al., 2025). This epistemic divergence is further evidenced by BNN uncertainty estimates, which were well calibrated for model predictions but inversely correlated with expert-perceived difficulty; where experts expressed uncertainty due to equifinality, missing context, or weak feature expression, the BNN remained confident because visual patterns matched training distributions. This confirms that Monte Carlo captures parameter uncertainty, not interpretive ambiguity (Gal and Ghahramani, 2016). Expert oversight therefor remains essential for low-confidence cases, non-diagnostic attention patterns as revealed through the expert-models comparison and interpretability analysis, especially when specific images challenge all classifiers (Bianchi et al., 2024; Steyvers et al., 2022).

However, several limitations constrain the generalizability of our findings. Our simplified classification scheme does not capture the full spectrum of frost action variation. The five CNN architectures evaluated here represent a subset of available deep learning approaches, and other architectures designed for spatial relationship capture may better accommodate the relational reasoning central to microfeatures. Reliance on PPL at single magnification eliminates information from XPL examination and multi-scale observation central to standard micromorphological practice.

While our dataset encompasses eleven Plio-Pleistocene sites, broader testing across diverse periglacial and non-periglacial contexts is needed to establish robust generalization. In particular, validation on fully independent datasets from sites not represented in the training data would be a critical next step for establishing model transferability across different sedimentary contexts, depositional environments, and imaging setups.

Future work should expand datasets across more sediment types and depositional environments, incorporate multiple imaging modalities, and explore more model architectures suited for capturing spatial relationships between micro-features. A promising direction would be the development of multi-scale architectures that explicitly integrate both local patch-level learning, capturing the diagnostic criteria such as aggregate geometry and frost-related porosity, as well as photomicrograph-level context, enabling feature localization across scales. In this study, we took a preliminary step in this direction by evaluating model performance on both cropped patches and photomicrographs, demonstrating that patch level training captures diagnostic detail in the broader-photomicrograph spatial scale. Building on this, future work could develop models that combine both scales (or more) in a unified framework alongside with expert-annotated diagnostic regions as additional training signals. This would directly constrain models’ attention to micromorphologically meaningful features rather than spurious image regularities.

Beyond architectural and dataset considerations, future work will investigate methodological modifications to the training pipeline. As datasets increase in size and are more diverse, systematic comparisons of random and strict crop-level and photomicrograph-level splitting strategies would provide insights into the conditions in which each approach achieves the best balance between sample size, generalizability, and data independence. This comparison allows for establishing best practices for dataset preparation in micromorphological deep learning workflows.

Additionally, the low inter-expert agreement observed in our survey suggests that single label annotations may not fully capture the interpretive variability inherent to micromorphological classification. The survey itself is subject to several constraints that limit its broader generalizability. The sample size (n =19) and uneven distribution of self-reported experience levels with the majority reporting intermediate frost-specific experience, reflect the specialized and relatively small nature of practicing micromorphologists, aspects that dictated our deliberately descriptive and exploratory treatment of the survey results throughout this study. Our survey was also based on voluntary recruitment, which further introduces potential self-selection bias toward micromorphologists familiar with computational approaches or interested in machine learning integration. Together these limitations suggest that the survey findings are better understood as an exploratory baseline rather than a representative benchmark of micromorphological practice broadly. Future work should incorporate multi-annotator protocols that preserve disagreement as information rather than resolving it into single labels, thus enabling uncertainty-aware training strategies that reflect genuine diagnostic ambiguity (Sheng and Zhang, 2019).

To facilitate the adoption and reproducibility of this framework, we provide a publicly accessible classification tool that implements the two-stage classification pipeline with interpretability maps, enabling researchers to obtain frost feature predictions for their own thin-section images with full transparency into model attention patterns (available at: https://huggingface.co/spaces/skkou/frost_classifier). However, the outputs should be treated as a first-pass suggestion and not as definitive micromorphological diagnoses and must be checked with established criteria and site context.

Conclusions

This study evaluated whether deep learning models can identify and classify frost-related micromorphological features at an expert level while maintaining interpretability aligned with established diagnostic criteria. We trained five CNN models on photomicrographs from multiple archaeological sites and validated them against both interpretability analysis and an expert survey.

Three principal findings emerge. First, off-the-shelf models achieve better classification performance, but interpretability analysis revealed reliance on spurious correlations rather than diagnostic features, demonstrating that good performance alone is not sufficient for evaluating models in micromorphology. Second, expert agreement on the same task was low, confirming that inter-observer variability is intrinsic to frost feature identification rather than a limitation specific to computational approaches. Third, mode and expert errors were largely independent, indicating that CNNs and expert micromorphologists capture different and complementary aspects of frost feature recognition.

Together, these findings establish a proof-of-concept for model-expert collaboration that addresses persistent challenges in archaeological micromorphology: scaling comparative analysis beyond what the manual examination permits, improving reproducibility through consistent pattern recognition, and broadening analytical access for researchers without extensive frost-specific training. The proposed framework, combining interpretability validation, uncertainty quantification, multi-expert consensus labelling, and model-expert workflow, provides a transferable template for computational integration across disciplines facing similar challenges with subtle, overlapping features subject to inter-observer variability.

Expanding this framework to incorporate broader sedimentary contexts and multiple imaging modalities will be essential for moving towards computational micromorphological tools. An effective implementation requires integrated workflows, where models handle the first pass screening while experts provide targeted validation and contextual synthesis into behavioral narratives. This complementarity preserves interpretive integrity while enabling analyses at scales previously constrained by time and expertise availability.

Ethics and consent

No ethical approval was needed for this study. The survey was completely anonymous, and we used no identifiable information on the survey participants.

Data availability

All data underlying this study are available from the Open Science Framework (OSF):

OSF: Frost Action Classification. https://doi.org/10.17605/OSF.IO/TYDSM (Kouki, 2026a).

The project includes the following data:

- Image dataset used for classification
- Image classification results
- Survey results
- Expert-CNN results
- Supplementary tables and figures

Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC BY 4.0).

Extended data

Extended data are available from the Open Science Framework (OSF):

https://doi.org/10.17605/OSF.IO/TYDSM (Kouki, 2026a).

The project contains the following extended data:

- Additional statistical analysis of model performance
- Statistical analysis of expert survey data
- Expert-CNN comparison analysis
- Additional plots and tables
- Guidelines for using the interactive classification tool

Analysis code available from: https://github.com/SofiaKouki/Frost-Action-Classification.git

Archived analysis code at time of publication: https://doi.org/10.17605/OSF.IO/TYDSM (Kouki, 2026b).

License: CC0 1.0.

Software availability

Software available from: https://huggingface.co/spaces/skkou/frost_classifier

Source code available from: https://github.com/SofiaKouki/Frost-Action-Classification.git

Archived source code at time of publication: https://doi.org/10.17605/OSF.IO/TYDSM (Kouki, 2026b).

License: CC0 1.0.

Acknowledgements

We would like to thank William C. Murphee for his assistance with the annotations, as well as Carlos Duarte Simões for providing samples and contributing to the annotation process. We are especially grateful to Miguel Soares Remiseiro for his help with the annotations, discussions, and for testing the interactive tool. We are also deeply grateful to the archaeological teams and lead researchers who provided the samples used to train our models, including Daniel S. Adler,

David Lordkipanidze, Nicholas Tushabramishvili, Shannon McPherron, Paul Goldberg, Dennis Sandgathe, Jonathan A. Haws, Tsenka R. Tsanova, David Braun, Carolina Mallol, Marie Soressi, Pablo Arias, Roberto Ontañón Peredo, Esteban Álvarez Fernández and Jesús F. Jordá Pardo and all people participated in the site excavations. We also thank all the micromorphologists who kindly participated in the survey. We would like to thank Georgios Σ. Polymeris and Maria João Martins for their valuable comments during the preparation of the manuscript. We would also like to thank Marta Lorenzon for her support. Finally, we thank the members of the ERC MATRIX team: Carli Peters, Hussein Knabar, and Pedro Coxito for their comments throughout this work. No generative AI tools were used in the preparation of this manuscript.

References

Abadi M, Barham P, Chen J, *et al.*: **TensorFlow: A system for large-scale machine learning [WWW Document].** 2016. (accessed 6.9.25).
Reference Source

Abid A, Abdalla A, Abid A, *et al.*: **Gradio: Hassle-Free Sharing and Testing of ML Models in the Wild.** 2019.
Publisher Full Text

Adebayo J, Gilmer J, Muelly M, *et al.*: **Sanity Checks for Saliency Maps.** 2020.
Publisher Full Text

Aldeias V, Goldberg P, Dibble HL, *et al.*: **Deciphering site formation processes through soil micromorphology at Contrebandiers Cave.** *Morocco. J. Hum. Evol.* 2014; **69**: 8–30.
PubMed Abstract | **Publisher Full Text**

Aldeias V, Sandgathe D, McPherron SJP, *et al.*: **Site Formation Histories and Context of Human Occupations at the Paleolithic Site of La Ferrassie (Dordogne, France).** *J. Paleolit. Archaeol.* 2023; **6**: 30.
Publisher Full Text

Arnay R, García-Villa P, Hernández-Aceituno J, *et al.*: **Improving micromorphological analysis with CNN-based segmentation of flint/obsidian, bone and charcoal.** *PLOS ONE.* 2026; **21**: e0340353.
Publisher Full Text

Arnay R, Hernández-Aceituno J, Mallol C: **Soil micromorphological image classification using deep learning: The porosity parameter.** *Appl. Soft Comput.* 2021; **102**: 107093.
Publisher Full Text

Bai K, Zhang Z, Jin S, *et al.*: **Rock image classification based on improved EfficientNet.** *Sci. Rep.* 2025; **15**: 18683.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, *et al.*: **Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI.** *Inf. Fusion.* 2020; **58**: 82–115.
Publisher Full Text

Basso M, Souza JP d P, Chinelatto GF, *et al.*: **Petrographic image classification of complex carbonate rocks from the Brazilian pre-salt using convolutional neural networks.** *Sci. Rep.* 2025; **15**: 30700.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Bergen KJ, Johnson PA, de Hoop MV, *et al.*: **Machine learning for data-driven discovery in solid Earth geoscience.** *Science.* 2019; **363**: eaau0323.
PubMed Abstract | **Publisher Full Text**

Bertran P, Andrieux E, Antoine P, *et al.*: **Pleistocene Involutions and Patterned Ground in France: Examples and Analysis Using a GIS Database.** *Permafr. Periglac. Process.* 2017; **28**: 710–725.
Publisher Full Text

Bertran P, Beauval C, Boulogne S, *et al.*: **Experimental archaeology in a mid-latitude periglacial context: insight into site formation and taphonomic processes.** *J. Archaeol. Sci.* 2015; **57**: 283–301.
Publisher Full Text

Bertran P, Klaric L, Lenoble A, *et al.*: **The impact of periglacial processes on Palaeolithic sites: The case of sorted patterned grounds.** *Quat. Int.* 2010; **1**: 17–29.
Publisher Full Text

Bianchi M, Santis AD, Tocchetti A, *et al.*: **Interpretable Network Visualizations: A Human-in-the-Loop Approach for Post-hoc Explainability of CNN-based Image Classification.** *Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence.* 2024; pp. 3698–3706.
Publisher Full Text

Bisong E: **Google Colaboratory.** Bisong E, editor. *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners.* Berkeley, CA: Apress; 2019; pp. 59–64.
Publisher Full Text

Cai Y-W, Qiu K-F, Petrelli M, *et al.*: **The application of "transfer learning" in optical microscopy: The petrographic classification of opaque minerals.** *Am. Mineral.* 2024; **109**: 2060–2072.
Publisher Full Text

Chollet F: **keras.** 2015.

Conover WJ: *Practical Nonparametric Statistics* Wiley; 3rd Edition. 1999.

Courty MA: *Soils and micromorphology in archaeology.* Cambridge: Cambridge University Press; 1989.

Courty M-A, Vallverdu J: **The microstratigraphic record of abrupt climate changes in cave sediments of the Western Mediterranean.** *Geoarchaeology.* 2001; **16**: 467–499.
Publisher Full Text

Dalhat MA, Osman SA: **Deep learning based identification of rock minerals from un-processed digital microscopic images of undisturbed broken-surfaces.** *Artif. Intell. Geosci.* 2025; **6**: 100127.
Publisher Full Text

Davies W, Van Andel TH, Weninger B: *The Human Presence in Europe during the Last Glacial Period I: Human Migrations and the Changing Climate.* Van Andel TH, Davies W, editors. McDonald Institute for Archaeological Research; 2003; pp. 31–56.

Dawson HL, Dubrule O, John CM: **Impact of dataset size and convolutional neural network architecture on transfer learning for carbonate rock classification.** *Comput. Geosci.* 2023; **171**: 105284.
Publisher Full Text

Deeben J, Hiddink H, Huisman DJ, *et al.*: **Middle Palaeolithic artefact migration due to periglacial processes; a geological investigation into near-surface occurrence of Palaeolithic artefacts (Limburg-Eastern Brabant coversand region, the Netherlands).** *Neth. J. Geosci.* 2010; **89**: 35–50.
Publisher Full Text

Demšar J: **Statistical comparisons of classifiers over multiple data sets.** *J. Mach. Learn. Res. JMLR.* 2006a; **7**: 1–30.

Dibble HL, Abodolahzadeh A, Aldeias V, *et al.*: **How did hominins adapt to Ice Age Europe without fire?.** 2017; **58**: S278–S287.
Publisher Full Text

Dvijotham K(D), Winkens J, Barsbey M, *et al.*: **Enhancing the reliability and accuracy of AI-enabled diagnosis via complementarity-driven deferral to clinicians.** *Nat. Med.* 2023; **29**: 1814–1820.
PubMed Abstract | **Publisher Full Text**

Ebert-Uphoff I, Hilburn KA: **Evaluation, Tuning and Interpretation of Neural Networks for Meteorological Applications.** 2020.
Publisher Full Text

Faria EL, Barbosa R, Coelho JM, *et al.*: **Automatic classification of Carbonatic thin sections by computer vision techniques and one-vs-all models.** *Artif. Intell. Geosci.* 2025; **6**: 100117.
Publisher Full Text

Gal Y, Ghahramani Z: **Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning.** *Proceedings of The 33rd International Conference on Machine Learning. Presented at the International Conference on Machine Learning.* PMLR; 2016; pp. 1050–1059.

Gavrilov AD, Jordache A, Vasdani M, *et al.*: **Preventing Model Overfitting and Underfitting in Convolutional Neural Networks.** *Int. J. Softw. Sci. Comput. Intell.* 2018; **10**: 19–28.
Publisher Full Text

Geirhos R, Jacobsen J-H, Michaelis C, *et al.*: **Shortcut Learning in Deep Neural Networks.** *Nat. Mach. Intell.* 2020; **2**: 665–673.
Publisher Full Text

Ghojogh B, Crowley M: **The Theory Behind Overfitting, Cross Validation, Regularization, Bagging, and Boosting: Tutorial.** 2023.
Publisher Full Text

Goldberg P, Aldeias V: **Why does (archaeological) micromorphology have such little traction in (geo)archaeology?.** *Archaeol. Anthropol. Sci.* 2018; **10**: 269–278.
Publisher Full Text

Goldberg P, Macphail RI, Carey C, *et al.*: *Practical and Theoretical Geoarchaeology.* Oxford: Wiley-Blackwell; 2nd ed. 2022.

Gorriz JM, Clemente RM, Segovia F, *et al.*: **Is K-fold cross validation the best model selection method for Machine Learning.** 2024.
Publisher Full Text

He K, Zhang X, Ren S, *et al.*: **Deep Residual Learning for Image Recognition.** 2015.
Publisher Full Text

Hemmer P, Schemmer M, Kühl N, *et al.*: **Complementarity in human-AI collaboration: concept, sources, and evidence.** *Eur. J. Inf. Syst.* 2025; **34**: 979–1002.
Publisher Full Text

Holcomb JA, McDonough KN, Rosencrance RL, *et al.*: **Frost Action during the Younger Dryas Inferred from Soil Micromorphology at Connley Cave 5, Oregon.** *PaleoAmerica.* 2023; **9**: 289–303.
Publisher Full Text

Iyas MR, Setiawan NI, Warmada IW: **Mask R-CNN for rock-forming minerals identification on petrography, case study at Monterado, West Kalimantan.** *E3S Web Conf.* 2020; **200**: 06007.
Publisher Full Text

Jiang S, Sweet L, Blougouras G, *et al.*: **How Interpretable Machine Learning Can Benefit Process Understanding in the Geosciences. Earths.** *Future.* 2024; **12**: e2024EF004540.
Publisher Full Text

Jospin LV, Buntine W, Boussaid F, *et al.*: **Hands-on Bayesian Neural Networks -- a Tutorial for Deep Learning Users.** *IEEE Comput. Intell. Mag.* 2022; **17**: 29–48.
Publisher Full Text

Karkanas P, Goldberg P: *Reconstructing Archaeological Sites: Understanding the Geoarchaeological Matrix.* John Wiley & Sons Ltd.; 2018.
Publisher Full Text

Kehl M, Marcazzan D, Miller CE, *et al.*: **The Upper Sedimentary Sequence of Grotta di Fumane, Northern Italy: A Micromorphological Approach to Study Imprints of Human Occupation and Paleoclimate Change.** *Geoarchaeology.* 2025; **40**: e70000.
Publisher Full Text

Kouki S: Frost action classification. [Dataset]. *OSF.* 2026a.
Publisher Full Text

Kouki S: **Frost action classification [Software].** *OSF.* 2026b.
Publisher Full Text

Lakshmanan V, Görner M, Gillard R: *Practical Machine Learning for Computer Vision.* O'Relly; 2021.

Landis JR, Koch GG: **The measurement of observer agreement for categorical data.** *Biometrics.* 1977; **33**: 159–174.
Publisher Full Text

Lipton ZC: **The Mythos of Model Interpretability.** 2017.
Publisher Full Text

Liu X, Jiang S, Wu R, *et al.*: **Automatic taxonomic identification based on the Fossil Image Dataset (>415,000 images) and deep convolutional neural networks.** *Paleobiology.* 2023; **49**: 1–22.
Publisher Full Text

Lundberg S, Lee S-I: **A Unified Approach to Interpreting Model Predictions.** 2017.
Publisher Full Text

Maitre J, Bouchard K, Bédard LP: **Mineral grains recognition using computer vision and machine learning.** *Comput. Geosci.* 2019; **130**: 84–93.
Publisher Full Text

Masson B: **Combustion Features and Periglacial Structures: a New Taphonomic Analysis of Mousterian Combustion Features at Saint-Vaast-la-Hougue (50).** *Palethnol. Archéologie Sci. Hum.* 2010.
Publisher Full Text

McGovern A, Lagerquist R, Gagne DJ, *et al.*: **Making the Black Box More Transparent: Understanding the Physical Implications of Machine Learning.** *Bull. Am. Meteorol. Soc.* 2019; **100**: 2175–2199.
Publisher Full Text

McHugh ML: **Interrater reliability: the kappa statistic.** *Biochem. Med.* 2012; **22**: 276–282.
Publisher Full Text

Minh APT: **Overview of Class Activation Maps for Visualization Explainability.** 2023.
Publisher Full Text

Moral PD, Nowaczyk S, Pashami S: **Why Is Multiclass Classification Hard?** *IEEE Access.* 2022; **10**: 80448–80462.
Publisher Full Text

Mukaka M: **A guide to appropriate use of Correlation coefficient in medical research. Malawi Med.** *J. J. Med. Assoc. Malawi.* 2012; **24**: 69–71.

Müller R, Kornblith S, Hinton GE: **When does label smoothing help?.** *Advances in Neural Information Processing Systems.* Curran Associates, Inc.; 2019.

Murphree WC, Ferro-Vázquez C, Kulakovska L, *et al.*: **Fire Use During the Last Glacial Maximum: Evidence From the Epigravettian at Korman' 9, Middle Dniester Valley, Ukraine.** *Geoarchaeology.* 2025; **40**: e70006.
Publisher Full Text

Nicosia C, Stoops G: *Archaeological Soil and Sediment Micromorphology.* John Wiley & Sons Ltd.; 2017.
Publisher Full Text

Pedregosa F, Varoquaux G, Gramfort A, *et al.*: **Scikit-learn: Machine Learning in Python.** *J. Mach. Learn. Res.* 2011; **12**: 2825–2830.

Petrelli M: **Machine Learning in Petrology: State-of-the-Art and Future Perspectives.** *Journal of Petrology|Oxford Academic. J. Petrol.* 2024; **65**.
Publisher Full Text

Phillips E, Lipka E, van der Meer JJM: **Micromorphological evidence of liquefaction, injection and sediment deposition during basal sliding of glaciers.** *Quat. Sci. Rev.* 2013; **81**: 114–137.
Publisher Full Text

Pires de Lima R, Duarte D, Nicholson C, *et al.*: **Petrographic microfacies classification with deep convolutional neural networks.** *Comput. Geosci.* 2020; **142**: 104481.
Publisher Full Text

Polat Ö, Polat A, Ekici T: **Automatic classification of volcanic rocks from thin section images using transfer learning networks.** *Neural Comput. & Applic.* 2021; **33**: 11531–11540.
Publisher Full Text

Raghu M, Zhang C, Kleinberg J, *et al.*: **Transfusion: Understanding Transfer Learning for Medical Imaging.** 2019.
Publisher Full Text

Rainio O, Teuho J, Klén R: **Evaluation metrics and statistical tests for machine learning.** *Sci. Rep.* 2024; **14**: 6086.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Raschka S: **Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning.** 2020.
Publisher Full Text

Reichstein M, Camps-Valls G, Stevens B, *et al.*: **Deep learning and process understanding for data-driven Earth system science.** *Nature.* 2019; **566**: 195–204.
PubMed Abstract | **Publisher Full Text**

Romero IC, Kong S, Fowlkes CC, *et al.*: **Improving the taxonomy of fossil pollen using convolutional neural networks and superresolution microscopy.** *Proc. Natl. Acad. Sci. USA.* 2020a; **117**: 28496–28505.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Romero M, Interian Y, Solberg T, *et al.*: **Targeted transfer learning to improve performance in small medical physics datasets.** *Med. Phys.* 2020b; **47**: 6246–6256.
PubMed Abstract | **Publisher Full Text**

Rosenfeld A, Zemel R, Tsotsos JK: **The Elephant in the Room.** 2018.
Publisher Full Text

Rubo RA, de Carvalho Carneiro C, Michelon MF, *et al.*: **Digital petrography: Mineralogy and porosity identification using machine learning algorithms in petrographic thin section images.** *J. Pet. Sci. Eng.* 2019; **183**: 106382.
Publisher Full Text

Rudin C: **Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead.** 2019.
Publisher Full Text

Şahin E, Arslan NN, Özdemir D: **Unlocking the black box: an in-depth review on interpretability, explainability, and reliability in deep learning.** *Neural Comput. & Applic.* 2025; **37**: 859–965.
Publisher Full Text

Sandler M, Howard A, Zhu M, *et al.*: **MobileNetV2: Inverted Residuals and Linear Bottlenecks.** 2019.
Publisher Full Text

Santos CFGD, Papa JP: **Avoiding Overfitting: A Survey on Regularization Methods for Convolutional Neural Networks.** *ACM Comput. Surv.* 2022; **54**: 213:1–213:25.
Publisher Full Text

Seabold S, Perktold J: **Statsmodels: Econometric and Statistical Modeling with Python.** *Presented at the 9th Python in Science Conference.* Austin, Texas: 2010; pp. 92–96.
Publisher Full Text

Seo W, Kim Y, Sim H, *et al.*: **Classification of igneous rocks from petrographic thin section images using convolutional neural network.** *Earth Sci. Inf.* 2022; **15**: 1297–1307.
Publisher Full Text

Shahack-Gross R: **Archaeological Micromorphology Self-Evaluation Exercise.** *Geoarchaeology.* 2016; **31**: 49–57.
Publisher Full Text

Sheng VS, Zhang J: **Machine Learning with Crowdsourcing: A Brief Summary of the Past Research and Future Directions.** *Proc. AAAI Conf. Artif. Intell.* 2019; **33**: 9837–9843.
Publisher Full Text

Spelmen VS, Porkodi R: **A Review on Handling Imbalanced Data, in: 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT).** *Presented at the 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT).* 2018; pp. 1–11.
Publisher Full Text

Stadlhofer A, Mezhuyev V: **Approach to provide interpretability in machine learning models for image classification.** *Ind. Artif. Intell.* 2023; **1**: 10.
Publisher Full Text

Steyvers M, Tejeda H, Kerrigan G, *et al.*: **Bayesian modeling of human–AI complementarity.** *Proc. Natl. Acad. Sci.* 2022; **119**: e2111547119.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Stoops G, editor. **Guidelines for Analysis and Description of Soil and Regolith Thin Sections.** *ASA, CSSA, and SSSA Books.* Madison, WI, USA: John Wiley & Sons, Ltd; 2nd ed. 2021.
Publisher Full Text

Straus LG: **Southwestern Europe at the Last Glacial Maximum.** *Curr. Anthropol.* 1991; **32**: 189–199.
Publisher Full Text

Su C, Xu S, Zhu K, *et al.*: **Rock classification in petrographic thin section images based on concatenated convolutional neural networks.** *Earth Sci. Inf.* 2020; **13**: 1477–1484.
Publisher Full Text

Tan M, Le QV: **EfficientNetV2: Smaller Models and Faster Training.** 2021.
Publisher Full Text

Tang DG, Milliken KL, Spikes KT: **Machine learning for point counting and segmentation of arenite in thin section.** *Mar. Pet. Geol.* 2020; **120**: 104518.
Publisher Full Text

Texier JP, Bertran P, Coutard JP, *et al.*: **TRANSIT, an experimental archaeological program in periglacial environment: Problem, methodology, first results.** *Geoarchaeology.* 1998; **13**: 433–473.
Publisher Full Text

Todisco D, Monchot H: **Bone Weathering in a Periglacial Environment: The Tayara Site (KbFk-7), Qikirtaq Island, Nunavik (Canada).** *Arctic.* 2008; **61**: 87–101.
Publisher Full Text

Tohka J, van Gils M: **Evaluation of machine learning algorithms for health and wellness applications: A tutorial.** *Comput. Biol. Med.* 2021; **132**: 104324.
PubMed Abstract | **Publisher Full Text**

Vachula RS, Huang Y, Longo WM, *et al.*: **Evidence of Ice Age humans in eastern Beringia suggests early migration to North America.** *Quat. Sci. Rev.* 2019; **205**: 35–44.
Publisher Full Text

Van Vliet-Lanoë B: **Frost and soils: implications for paleosols, paleoclimates and stratigraphy.** *Catena.* 1998; **34**: 157–183.
Publisher Full Text

Van Vliet-Lanoë B: **Frost effects in soils.** *Soil and Quaternary Landscape Evolution.* John Wiley & Sons Ltd; 1985; pp. 115–156.

Van Vliet-Lanoë B, Fox CA, Gubin SV: **Micromorphology of Cryosols.** Kimble JM, editor. *Cryosols: Permafrost-Affected Soils.* Berlin, Heidelberg: Springer; 2004; pp. 365–390.
Publisher Full Text

Vliet-Lanoë BV: **6 - Frost Action.** Stoops G, Marcelino V, Mees F, editors. *Interpretation of Micromorphological Features of Soils and Regoliths.* Amsterdam: Elsevier; 2010; pp. 81–108.
Publisher Full Text

Wang H, Wang Z, Du M, *et al.*: **Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks.** 2020.
Publisher Full Text

Wang H, Yeung D-Y: **A Survey on Bayesian Deep Learning.** 2021.
Publisher Full Text

Ying X: **An Overview of Overfitting and its Solutions.** *J. Phys. Conf. Ser.* 2019; **1168**: 022022.
Publisher Full Text

Zhang J, Chao H, Dasegowda G, *et al.*: **Overlooked Trustworthiness of Saliency Maps.** Wang L, Dou Q, Fletcher PT, *et al.*, editors. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022.* Switzerland, Cham: Springer Nature; 2022; pp. 451–461.
Publisher Full Text

Zhao Z, Alzubaidi L, Zhang J, *et al.*: **A comparison review of transfer learning and self-supervised learning: Definitions, applications, advantages and limitations.** *Expert Syst. Appl.* 2024; **242**: 122807.
Publisher Full Text

Zheng D, Zhong H, Camps-Valls G, *et al.*: **Explainable deep learning for automatic rock classification.** *Comput. Geosci.* 2024; **184**: 105511.
Publisher Full Text

Zickel M, Gröbner M, Röpke A, *et al.*: **MiGIS: micromorphological soil and sediment thin section analysis using an open-source GIS and machine learning approach.** *EG Quat. Sci. J.* 2024; **73**: 69–93.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ✔ ✔ ? ✔

Version 2

Reviewer Report 17 September 2026

https://doi.org/10.21956/openreseurope.26759.r79362

© **2026 Hein A.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

✔ **Anno Hein**
Institute of Nanoscience and Nanotechnology, National Centre of Scientific Research "Demokritos", Aghia Paraskevi, Attica, Greece

The present manuscript concerns the examination and categorization of petrographic thin-sections of soil samples with the scope to investigate their exposition to freeze thaw cycles. For this, different models for supervised machine learning are tested and compared with classifications of a common set of micrographs by experts for micro-morphological studies. I was not involved in the reviewing process of the first version. Nevertheless, having read the comments and suggestions expressed in the reviews of the initial manuscript and the changes and replies by the authors I acknowledge that the reviews were appropriately taken into consideration.

Is the work original in terms of material and argument?
Yes

Does it sufficiently engage with relevant methodologies and secondary literature on the topic?
Yes

Is the work clearly and cogently presented?
Yes

Is the argument persuasive and supported by evidence?
Yes

If any, are all the source data and materials underlying the results available?
Yes

Does the research article contribute to the cultural, historical, social understanding of the field?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Analytical studies of ancient and traditional ceramics including micro-structural classification; Investigation of deterioration processes of Cultural Heritage materials

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 22 August 2026

https://doi.org/10.21956/openreseurope.26759.r78563

© 2026 Bellat M. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Mathias Bellat
University of Tübingen, Tübingen, Germany

All the concerns expressed earlier have been addressed. I therefore would recommend indexing.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Archaeological sciences (machine learning); Southwestern Asia Neolithic archaeology; Landscape archaeology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 19 August 2026

https://doi.org/10.21956/openreseurope.26759.r78562

© 2026 Gacu J. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Jerome G. Gacu
Kyoto University, Kyoto, Japan

All comments were resolved.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Sediment management, water resources engineering

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 15 June 2026

https://doi.org/10.21956/openreseurope.25655.r73575

© 2026 Karamitrou A. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Alexandra Karamitrou
University of Southampton, Southampton, UK

Structure
Although the supplementary material contains substantial detail, the manuscript itself needs additional explanation and visual support to be fully accessible to non-expert readers. Consider integrating key figures and clarifying complex sections within the main text.

Clarity
- In the *Machine Learning Workflow* section, please provide additional detail on each method used, including relevant citations. This will help readers understand the rationale behind your choices and how each component contributes to the workflow.

- The *Experts and CNN Models Alignment* section would be clearer if the results were summarized in a table. This would allow you to remove the dense explanatory text at the end of the section and present the findings more effectively.

Figures

-The *Materials and Methods* section would benefit from including example photomicrographs, as well as a map showing the locations of all sampled materials. These additions would greatly improve clarity for readers unfamiliar with the dataset.

- Figure S.4.2.1 from the supplementary material should be incorporated into the main manuscript. It provides a clear visual representation of the workflow, which is difficult to follow when described only in text.

- While the supplementary material is extensive, the manuscript itself would benefit from relocating some of these figures into the main text to support non-expert readers.

Technical Corrections

-“All the images we acquired are in RGB format, with a resolution of 2880 x 2048-pixel resolution and saved as uncompressed.tif files.”
Correct to:
“All the images we acquired are in RGB format, with a resolution of 2880 x 2048-pixel and saved as uncompressed.tif files.”

Supplementary material

I can verify that the code runs correctly and that all necessary documents are accessible through the provided links.

Is the work original in terms of material and argument?
Yes

Does it sufficiently engage with relevant methodologies and secondary literature on the topic?
Yes

Is the work clearly and cogently presented?
Partly

Is the argument persuasive and supported by evidence?
Yes

If any, are all the source data and materials underlying the results available?
Yes

Does the research article contribute to the cultural, historical, social understanding of the field?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: My area of research is in the use of artificial intelligence and remote sensing data (or images) to detect and map different types of archaeological targets or targets of archaeological interest.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 23 Jul 2026

Sofia Kouki

We sincerely thank the reviewer for their constructive and insightful feedback, as well as for

the time and effort dedicated to reviewing our manuscript. Below, we address each comment point-by-point. Structure Although the supplementary material contains substantial detail, the manuscript itself needs additional explanation and visual support to be fully accessible to non-expert readers. Consider integrating key figures and clarifying complex sections within the main text. We have now added additional figures and tables from the Supplementary Information to the manuscript.
Specifically, the new additions include the workflow image (Figure 2), the confusion matrices of the misclassifications (now as Figure 4a), and additional images of the model-expert comparison (now as Figure 8). Clarity - In the Machine Learning Workflow section, please provide additional detail on each method used, including relevant citations. This will help readers understand the rationale behind your choices and how each component contributes to the workflow.
We revised parts of the Machine Learning workflow and provided the rationale behind each step and how each one of them contributes to the workflow (lines: 136-163, 155-158, and 162-168) - The Experts and CNN Models Alignment section would be clearer if the results were summarized in a table. This would allow you to remove the dense explanatory text at the end of the section and present the findings more effectively. We revised the Expert and CNN model Alignment section, and we added a table and a figure (now Table 2 and Figure 8) to facilitate non-expert readers.
Figures -The Materials and Methods section would benefit from including example photomicrographs, as well as a map showing the locations of all sampled materials. These additions would greatly improve clarity for readers unfamiliar with the dataset. We included a figure that contains a map with the sites used for this study and an example of photomicrographs from the locations (now Figure 1).
We also created an interactive map of the sites that contain representative photomicrographs from the sites, and we included it in the OSF repository so it can be publicly accessible (10.17605/OSF.IO/TYDSM). - Figure S.4.2.1 from the supplementary material should be incorporated into the main manuscript. It provides a clear visual representation of the workflow, which is difficult to follow when described only in text.
Figure S4.2.1 from the Supplementary Information is now incorporated into the main text as Figure 2. - While the supplementary material is extensive, the manuscript itself would benefit from relocating some of these figures into the main text to support non-expert readers.
We included additional figures and tables from the Supplementary Information in the main text to support non-expert readers (now as Figure 2 and 8 and Table 2). Technical Corrections -“All the images we acquired are in RGB format, with a resolution of 2880 x 2048-pixel resolution and saved as uncompressed.tif files.” Correct to: “All the images we acquired are in RGB format, with a resolution of 2880 x 2048-pixel and saved as uncompressed.tif files.” Corrected. Supplementary material I can verify that the code runs correctly and that all necessary documents are accessible through the provided links. We thank the reviewer for this verification.

Competing Interests: No competing interests were disclosed.

Reviewer Report 11 June 2026

https://doi.org/10.21956/openreseurope.25655.r75439

© **2026 Gacu J.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Jerome G. Gacu
Kyoto University, Kyoto, Japan

1. Add a methodology framework figure in the main manuscript showing image acquisition, preprocessing, data splitting, CNN training, Score-CAM analysis, expert survey, and model validation workflow.
2. Redesign Figure 1 using a common x-axis and shared legend. Increase subplot size to improve readability and utilize page space more efficiently.
3. In Figure 2, add arrows, circles, or labels to identify the specific sediment features causing model misclassification. The current examples are difficult to interpret without annotations.
4. Figure 3, use a single shared color bar and legend for all Score-CAM outputs. Increase image size and annotate representative regions showing spurious attention and diagnostically relevant features.
5. Perform a strict photomicrograph-level train-validation-test split and compare results with the current crop-level split. Report performance differences in the main manuscript.
6. Add a table summarizing final hyperparameters for each model, including learning rate, batch size, dropout rate, optimizer, epochs, and regularization settings.
7. If possible, validate the best-performing model using an independent dataset from a different archaeological site to demonstrate model generalizability.
8. Simplify Figure 5c or divide it into two figures. The current figure is visually crowded and difficult to interpret.
9. Include 95% confidence intervals for all reported performance metrics (Accuracy, Precision, Recall, and F1-score).
10. Add a discussion on the limitations of the expert survey, particularly the small sample size and uneven distribution of expertise levels.

Is the work original in terms of material and argument?
Partly

Does it sufficiently engage with relevant methodologies and secondary literature on the topic?
Yes

Is the work clearly and cogently presented?
Partly

Is the argument persuasive and supported by evidence?
Partly

If any, are all the source data and materials underlying the results available?
No source data required

Does the research article contribute to the cultural, historical, social understanding of the field?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Sediment management, water resources engineering

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 23 Jul 2026

Sofia Kouki

We would like to thank the reviewer for reading our manuscript and providing meaningful comments that improved our draft. We appreciate the time and effort dedicated to reviewing our work. Below, we address each comment in turn.
1. Add a methodology framework figure in the main manuscript showing image acquisition, preprocessing, data splitting, CNN training, Score-CAM analysis, expert survey, and model validation workflow. We added a workflow image to the main draft, now as Figure 2.

2. Redesign Figure 1 using a common x-axis and shared legend. Increase subplot size to improve readability and utilize page space more efficiently. We have redesigned the figure (now as Figure 1) in response to this comment. Specifically, all subplots now have a shared x-axis across both rows, legends are consolidated into a single shared legend, and we increased the size to improve readability as suggested.

3. In Figure 2, add arrows, circles, or labels to identify the specific sediment features causing model misclassification. The current examples are difficult to interpret without annotations. We revised the image by adding arrows and lines, showing better the features that cause models' misclassifications.

4. Figure 3, use a single shared color bar and legend for all Score-CAM outputs. Increase image size and annotate representative regions showing spurious attention and diagnostically relevant features. We revised the figure as suggested; it is now Figure 5 in the manuscript.

5. Perform a strict photomicrograph-level train-validation-test split and compare results with the current crop-level split. Report performance differences in the main manuscript. We thank the reviewer for raising this point regarding dataset splitting, and we assume that this comment is raised because of the potential risk of performance inflation due to crop-level dependency. We agree that in microscopic imaging, executing data splits at the independent image or slide level is another way to test absolute model generalizability across sediments. We have added a paragraph (lines: 609-619) to discuss these data partitioning dynamics, and we acknowledge that the spatial autocorrelation bias is not fully

neutralized. However, we respectfully argue that for this specific scope, sample size constraints, and primary objectives of this study, our current crop-level splitting strategy is methodologically justified and robust. We have chosen not to repartition and retrain the models at the strict photomicrograph level for the following reasons: Targeted crop curation: In standard computer vision applications, crop-level data leakage occurs when an image is sliced into redundant, overlapping grid tiles that capture the exact same uniform background signal. In our workflow, crops were not arbitrary spatial subdivisions. Instead, they were deliberately curated to retain only regions with diagnostic features, while discarding non-diagnostic elements (e.g., limestone fragments and slide edges). Because of this strict semantic filter, crops originating from the same photomicrograph do not represent redundant samples of a singular background texture; rather, each retained crop captures a distinct diagnostic observation. This makes the crop-level more appropriate analytical unit for our task. Contextual diversity and sample-size constraints: Our models were tasked with generalizing across diverse archaeological sites with different depositional processes. Because diagnostic freeze-thaw features can be localized and do not distribute uniformly across the entire slide, a strict photomicrograph level split across the entire dataset would have restricted the network's exposure to minority feature variation per site. The crop-level split allowed us to maximize the model's learning capabilities and ensure stable convergence across the range of heterogeneous contexts. To verify the validity of our choice, we conducted an independent, whole photomicrograph-level robustness test (discussed in the manuscript in lines 194-201 and 376-378). As detailed in Supplementary Information 4, Section 4.6, this independent evaluation confirmed that the comparative architectural behaviors and interpretability trends reported remain robust. Nonetheless, we added in the Discussion an explicit mention for future integration of this slide-level split as the underlying dataset expands.

6. Add a table summarizing final hyperparameters for each model, including learning rate, batch size, dropout rate, optimizer, epochs, and regularization settings. A comprehensive summary of final hyperparameters for all five models is already included in Supplementary Information 4, Section S2.4, Table 4.2.2. In the manuscript there is an explicit reference to this table in the Materials and Methods section (Materials and Methods, Machine Learning Workflow, paragraph 4). We believe that retaining the hyperparameter table in the supplementary material is appropriate given the level of technical detail involved, which would otherwise interrupt the flow of the method section, while remaining fully accessible to readers who wish to reproduce or build upon our workflow.

7. If possible, validate the best-performing model using an independent dataset from a different archaeological site to demonstrate model generalizability. We fully agree with the reviewer that validation on an independent dataset from different archaeological sites would strengthen the generalizability claims of the current models, and we recognize this as a priority for future work. Unfortunately, at this stage we do not have access to additional independently collected thin section material from sites outside our current dataset that would constitute a truly independent validation set. However, we would like to draw the reviewer's attention to several steps we did toward generalizability within the constraints of our current dataset: (1) our dataset already spans 11 archaeological sites with diverse periglacial and non-periglacial depositional environments; (2) our robustness test on full photomicrographs provides an additional evaluation that is independent of the crop-level

splitting strategy; and (3) our held-out test set was kept entirely separate from training and validation throughout. To this end, we have added an explicit statement to the limitations and future work section of the discussion acknowledging independent dataset validation as a critical next step, alongside a broader sedimentary context and depositional environments that we discussed already (lines: 603-606).

8. Simplify Figure 5c or divide it into two figures. The current figure is visually crowded and difficult to interpret. We thank the reviewer for this suggestion. We acknowledge that in the previous version of the manuscript, Figure 5c (now Figure 7c) conveys a lot of information; however, this plot is designed to simultaneously convey per-image error rates for both experts and models, their directional relationship, and complementary patterns across all 14 classification images. We believe this format is the most efficient way to present these four dimensions of information in a single panel without losing both modeling and micromorphological details. The thumbnail photomicrographs are an intentional design choice that allows readers to directly connect classification difficulty patterns to the images.

9. Include 95% confidence intervals for all reported performance metrics (Accuracy, Precision, Recall, and F1-score). We have now included the 95% confidence intervals for all reported metrics in the figure (now as Figure 3). These changes were implemented alongside the figure redesign requested in Comment 2. We also updated the image legend and caption to include the CI. 10. Add a discussion on the limitations of the expert survey, particularly the small sample size and uneven distribution of expertise levels. We have expanded the discussion of the expert survey constraints to the limitations and future work section of the discussion (lines 627-641), where it sits alongside the other methodological limitations of the study. We believe that this placement is more consistent with the overall structure of the discussion while ensuring the limitations are clearly acknowledged as the reviewer proposed.

Competing Interests: No competing interests were disclosed.

Reviewer Report 02 June 2026

https://doi.org/10.21956/openreseurope.25655.r73580

© 2026 Bellat M. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Mathias Bellat
University of Tübingen, Tübingen, Germany

Relevance of the manuscript to the archaeological field:

This manuscript presents a new methodology for discerning frozen features in thin sections using an image recognition algorithm. By comparing different deep learning and other types of models

with experts' assessments, it distinguishes differences between frozen and non-frozen sections, and, for the frozen section, three types of classes. Results show better performance of two deep learning models compared to expert-expert agreement, while the models' attention is not drawn to the distinctive criteria.

Abstract and Keywords:

Background section: "Identifying these features through thin section micromorphology is well established" I don't see the meaning of "well established"; consider replacing it with something else, like "these features are possible to identify under a microscope ".
Methods section: "(first presence, then feature type)" Please consider adding the notion of *multiclass labelling* to distinguish these two approaches from the abstract onward.
Article

Major comments:
I think the switch from §1.4 to §1.5 in the introduction is a little bit too straightforward. You should briefly introduce the computational approaches before explaining WHY they are particularly suited for frozen action. In my opinion, your construction should be :
Frozen action description (§1.3) -> difficulty of identification of these features (§1.4 without last sentence) -> computational approaches may fill this "Gap" (reframe §1.5) -> explain principles and limitations of the computational approaches (§1.6 - §1.8).
This work should not take long; all the elements are already there and well explained, it is just a matter of logical organisation of the arguments, which might lose a non-specialist reader.

Site selection and data preparation:
Did you also have several experts attest to the identification of thin-section features, or was it done by only one researcher?

Minor comments:
Introduction:
"microenvironments", while I understand the idea, I do not entirely agree. Features revealing wetter conditions (vesicle, riozome) can be linked with larger climatic conditions, therefore not "micro". I think environments alone is sufficient and more adapted to all possible conditions.

Regarding the interpretability of the model, while your approach is entirely justified and correct, you might consider adding a few words on the SHAP value analysis, which also allows interpretability for image classification (https://shap.readthedocs.io/en/latest/image_examples.html).

Material and methods:

"In our case, a 5x magnification provides a balance between the feature resolution". While you justify the choice of this magnification above, are you not afraid of missing many feature observations? Rizhoms or desiccation can only be seen at higher magnification most of the time. Same for the silt and clay structure.

With ImageNet weights models, did you use a specific pre-trained version of the algorithms? In which case, you can cite the affiliated papers or GitHub repositories to enhance reproducibility.

"We evaluated all models' performance using accuracy, precision, recall, F1-scores, and confusion matrices, as well as visualising the misclassified images per mode". Could you provide the equations for each metric? Integrate the S4.1 info somehow into the appendix or in the text.

"Fleiss' kappa, individual expert-reference labels alignment using Cohen's kappa " *Idem* for these metrics

What is the demography of the experts? In the supplementary material, I could not find any information on the academic background (researcher, field archaeologist, technicians) or geographical location (country of affiliation) of the experts.

Workflow and interactive classification tool:
I just wanted to highlight the quality of the tool the authors produced. This truly enhanced the accessibility of the research to non-developers.

Results:
"Overall, all five models (MobileNetV2, EfficientNetV2B0, ResNet50, Custom CNN, and BNN) perform well on both binary (frost vs. non-frost) and multiclass (blocky, granular, and platy-lenticular) tasks in both 5-fold cross-validation and the test set. " Please do not use the element's **good result**; this does not make sense if you are not comparing your results to other models. Use the range of your score instead.

"Agreement between experts and model approaches the level of agreement observed between experts." I don't understand this sentence.

Discussion:
Could you think of a way – for future applications - to constrain a deep learning model to focus on diagnostic elements rather than just a patch?

Language
"Israel, Levant", while I do not agree with the current politics of Israel, I think you should remove the " Levant " mention for a site which is not considered part of the state of Palestine. If you have a specific issue with this mention, please contact the editors, as I cannot assess this kind of decision as a reviewer.

"uncompressed.tif " lacks a space between the word and the later dot.

Graphics, images and tables
Consider adding a map showing all site locations and a distinct sign for the site experiencing frozen or non-frozen conditions.
Legend of Fig.4, the "(0–80% scale)" word uses a double score instead of a single score.

Please consider adding the supplementary figure Figure S.4.2.1 to the core text.
I could not verify the code due to my platform system (ArchLinux)

Bibliography
The authors have provided all necessary references regarding the topic. I would suggest

additional references, however, as they might bring elements regarding the presented subject

Additional references:
Zickel, M., Gröbner, M., Röpke, A., & Kehl, M. (2024). MiGIS: Micromorphological soil and sediment thin section analysis using an open-source GIS and machine learning approach. *E&G Quaternary Science Journal*, *73*(1), 69–93. https://doi.org/10.5194/egqsj-73-69-2024

References
1. Zickel M, Gröbner M, Röpke A, Kehl M: MiGIS: micromorphological soil and sediment thin section analysis using an open-source GIS and machine learning approach. *E&G Quaternary Science Journal*. 2024; **73** (1): 69-93 Publisher Full Text

Is the work original in terms of material and argument?
Yes

Does it sufficiently engage with relevant methodologies and secondary literature on the topic?
Yes

Is the work clearly and cogently presented?
Yes

Is the argument persuasive and supported by evidence?
Yes

If any, are all the source data and materials underlying the results available?
Yes

Does the research article contribute to the cultural, historical, social understanding of the field?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Archaeological sciences (machine learning); Southwestern Asia Neolithic archaeology; Landscape archaeology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 23 Jul 2026

Sofia Kouki

We would like to thank the reviewer for reading our manuscript and the constructive comments. We appreciate the time and effort dedicated to evaluating our work. Below, we address each comment in turn.

Relevance of the manuscript to the archaeological field: This manuscript presents a new methodology for discerning frozen features in thin sections using an image recognition algorithm. By comparing different deep learning and other types of models with experts' assessments, it distinguishes differences between frozen and non-frozen sections, and, for the frozen section, three types of classes. Results show better performance of two deep learning models compared to expert-expert agreement, while the models' attention is not drawn to the distinctive criteria.

Abstract and Keywords: Background section: "Identifying these features through thin section micromorphology is well established" I don't see the meaning of "well established"; consider replacing it with something else, like "these features are possible to identify under a microscope ". This sentence was changed to read: "Frost-associated features can be identified through soil micromorphology techniques using well-defined diagnostic criteria". We chose to retain the term "well-defined" because it reflects not only the ability to examine soils and sediments under a microscope but also the long-established application of soil micromorphology to the study of frost-affected sites. Moreover, this technique preserves the original geometry of the deposits, allowing the investigation of pore geometries that would otherwise be impossible to identify. Methods section: "(first presence, then feature type)" Please consider adding the notion of multiclass labelling to distinguish these two approaches from the abstract onward. We changed the sentence to "...implementing a two-stage classification approach by detecting first the presence of those features (binary classification) and then the feature type (multiclass classification)," see Abstract Method section.

Article Major comments: I think the switch from §1.4 to §1.5 in the introduction is a little bit too straightforward. You should briefly introduce the computational approaches before explaining WHY they are particularly suited for frozen action. In my opinion, your construction should be: Frozen action description (§1.3) -> difficulty of identification of these features (§1.4 without last sentence) -> computational approaches may fill this "Gap" (reframe §1.5) -> explain principles and limitations of the computational approaches (§1.6 - §1.8). This work should not take long; all the elements are already there and well explained, it is just a matter of logical organization of the arguments, which might lose a non-specialist reader. We have reorganized the sections as requested (lines 41-48).

Site selection and data preparation: Did you also have several experts attest to the identification of thin-section features, or was it done by only one researcher? The rationale and the labeling procedure are described in Supplementary Information 4 (Section 4.2, Data labeling), and we expanded this section to contain extensive details about this topic.

Minor comments: Introduction: "microenvironments", while I understand the idea, I do not entirely agree. Features revealing wetter conditions (vesicle, riozome) can be linked with

larger climatic conditions, therefore not "micro". I think environments alone is sufficient and more adapted to all possible conditions. We agree with the reviewer, and we have changed the word from "microenvironments" to “environments” as suggested (line 9). Regarding the interpretability of the model, while your approach is entirely justified and correct, you might consider adding a few words on the SHAP value analysis, which also allows interpretability for image classification (https://shap.readthedocs.io/en/latest/image_examples.html). We thank the reviewer for this suggestion. We have revised the relevant sentence in the introduction to broaden the interpretability methods with an explicit acknowledgement of SHAP (lines: 55-58)

Material and methods: "In our case, a 5x magnification provides a balance between the feature resolution". While you justify the choice of this magnification above, are you not afraid of missing many feature observations? Rizhoms or desiccation can only be seen at higher magnification most of the time. Same for the silt and clay structure. We would like to clarify that rhizomorphs and larger desiccation features fall outside the scope of this study’s classification framework, which specifically targets aggregate-scale frost microstructures (granular, blocky, and platy-lenticular). These microstructures are defined by aggregate geometry, porosity distribution, and spatial relationships between features, which are all optimally captured at 5x magnification. Higher magnifications would resolve fine fabric details but would simultaneously lose the spatial contexts necessary to distinguish between the three target microstructure classes. The 5x magnification was therefore not a compromise, but a deliberate methodological choice driven by the classification objectives.

With ImageNet weights models, did you use a specific pre-trained version of the algorithms? In which case, you can cite the affiliated papers or GitHub repositories to enhance reproducibility. The ImageNet pre-trained weights for all three off-the-shelf architectures (MobileNetV2, EfficientNetV2B0, and ResNet50) were loaded through the standard Keras applications API, with versions determined by Tensorflow environment reported in the manuscript. This information is also contained in the requirements.txt file on GitHub where the code is publicly available (https://github.com/SofiaKouki/Frost-Action-Classification.git).

We added a clarifying sentence in lines 295-297 to showcase the pre-trained model with their original papers cited. "We evaluated all models' performance using accuracy, precision, recall, F1-scores, and confusion matrices, as well as visualising the misclassified images per mode". Could you provide the equations for each metric? Integrate the S4.1 info somehow into the appendix or in the text. "Fleiss’ kappa, individual expert-reference labels alignment using Cohen’s kappa " Idem for these metrics We have now expanded the existing glossary table in Supplementary Information 4 (Table S4.1.1) to include formal equations for all performance metrics and statistical measures used in this study, alongside their definitions and references.

We believe this provides a cleaner and more accessible reference for readers without adding unnecessary complexity to the text. We point to the Supplementary Information in the main text for further information on the metrics for the readers to look for (line: 280).

What is the demography of the experts? In the supplementary material, I could not find any information on the academic background (researcher, field archaeologist, technicians) or geographical location (country of affiliation) of the experts. This is a valid point, as we only collected self-reported frost experience levels as the most directly relevant proxy for experience, and we acknowledged the limited demographic characterization as a design limitation in the manuscript. However, there are two main reasons why we chose not to

collect demographic information on participants' academic background or geographical affiliation.
First, given the small sample size (n = 19) and the specialized nature of the micromorphological community, collecting such details would risk identifying individual participants, compromising the ethical integrity of the survey. Second, and more fundamentally, demographic variables play no role in our statistical analysis. The expertise level we report is used descriptively rather than as the basis for inferential conclusions and adding further demographic stratification to a sample of 19 participants would not yield statistically meaningful results. Still, we agree with the reviewer that more in-depth surveys would be of interest to assess variance in micromorphologists' identification of frost action as well as of other processes. This is an important avenue for future research but lies beyond the scope of the present study.

Workflow and interactive classification tool: I just wanted to highlight the quality of the tool the authors produced. This truly enhanced the accessibility of the research to non-developers. We thank the reviewer for the supportive comment.
Results: "Overall, all five models (MobileNetV2, EfficientNetV2B0, ResNet50, Custom CNN, and BNN) perform well on both binary (frost vs. non-frost) and multiclass (blocky, granular, and platy-lenticular) tasks in both 5-fold cross-validation and the test set. " Please do not use the element's good result; this does not make sense if you are not comparing your results to other models. Use the range of your score instead. We have replaced the evaluative descriptors with specific performance ranges for binary and multiclass classification tasks, respectively (lines: 290-293) "Agreement between experts and model approaches the level of agreement observed between experts." I don't understand this sentence. We have removed the unclear sentence and retained the more explicit reporting of pairwise agreement values, sample sizes, and statistical significance in the following sentence, which more clearly conveys that models achieve inter-rater reliability comparable to that of human experts (lines: 446-449).
Discussion: Could you think of a way – for future applications - to constrain a deep learning model to focus on diagnostic elements rather than just a patch? We agree with the reviewer that this is a meaningful future application. To express that, we have now added a full paragraph in this direction in the limitations and future work section of the Discussion (lines: 665-671).
Language "Israel, Levant", while I do not agree with the current politics of Israel, I think you should remove the " Levant " mention for a site which is not considered part of the state of Palestine. If you have a specific issue with this mention, please contact the editors, as I cannot assess this kind of decision as a reviewer. In the spirit of consistency, we have removed the "Levant," and we kept the countries' names. "uncompressed.tif " lacks a space between the word and the later dot. Corrected.
Graphics, images and tables Consider adding a map showing all site locations and a distinct sign for the site experiencing frozen or non-frozen conditions. We have added a map showing the site locations to the main draft, now as Figure 1. Legend of Fig.4, the "(0–80% scale)" word uses a double score instead of a single score. Corrected. Please consider adding the supplementary figure Figure S.4.2.1 to the core text. We have added the workflow image to the main draft, now as Figure 2. I could not verify the code due to my platform system (ArchLinux). The code is publicly accessible through the GitHub repository (https://github.com/SofiaKouki/Frost-Action-Classification.git) and archived on

OFS (10.17605/OSF.IO/TYDSM), along with the saved models. The entire workflow was implemented in Google Collaboratory, a cloud-based platform that runs independently of the user's local operating system and is accessible through any standard web browser. On the GitHub page, there are also guidelines on how to implement the pipeline without depending on platform constraints.
Bibliography The authors have provided all necessary references regarding the topic. I would suggest additional references, however, as they might bring elements regarding the presented subject Additional references: Zickel, M., Gröbner, M., Röpke, A., & Kehl, M. (2024). MiGIS: Micromorphological soil and sediment thin section analysis using an open-source GIS and machine learning approach. E&G Quaternary Science Journal, 73(1), 69–93. https://doi.org/10.5194/egqsj-73-69-2024 Zickel et al. (2024) differs methodologically from our approach, since the authors rely on traditional machine learning rather than deep learning, and operating on whole thin section scans. Nevertheless, in the interest of providing a comprehensive overview of existing computational approaches in thin section analysis through soil micromorphology, we have added Zickel et al. (2024) to the introduction along with the rest of the studies, noting its distinct scope and methodology (lines: 69-72).

Competing Interests: No competing interests were disclosed.