Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Demonstrations Track

SoilNet App: AI-Assisted Expert-level Annotations of Soil Horizons

Vipin Singh[1] , **Joey Pruessing**[1] , **Teodor Chiaburu**[1] , **Einar Eberhardt**[2] ,
Sina Hesse[2] , **Stefan Broda**[2] , **Frank Haußer**[1] and **Felix Biessmann**[1,3]
[1]Berlin University of Applied Sciences, Berlin, Germany
[2]Federal Institute for Geosciences and Natural Resources, Hannover, Germany
[3]Einstein Center Digital Future, Berlin, Germany
{vipin.singh, s85955, chiaburu.teodor, frank.hausser, felix.biessmann}@bht-berlin.de,
{einar.eberhardt, sina.hesse, stefan.broda}@bgr.de

Abstract

Precise descriptions of soil horizons are required for policy makers, agriculture and many applications in civil engineering. Up to date correct soil horizon annotations require human experts as they follow complex hierarchical taxonomies. We present the SoilNet App, a web-based demonstrator that guides experts through relevant tasks for expert-level soil horizon annotations from soil profile images. To demonstrate the reliability of the SoilNet app we present results of a user study with soil horizon annotation experts, which highlights the difficulty of image-only-based annotation and suggests that collaborating with our model not only increases expert performance but also improves inter-annotator consistency. Our app is publicly accessible (https://soilnet.demo.calgo-lab.de).

1 Introduction

Soils sustain ecosystems, agriculture, and water filtration while regulating carbon and supporting biodiversity, making soil condition a direct driver of resilience and well-being [FAO, 2015]. Monitoring soil condition at scale is therefore essential for societies and policy makers. Soil horizons, the characterization of the soil's vertical and lateral structure, are core units for soil classification and monitoring [Hartemink *et al.*, 2020], captured in international and national systems such as WRB [IUSS Working Group WRB, 2022] and KA5 [Ad-hoc-AG Boden, 2005]. Producing high-quality horizon labels is labor-intensive, requires trained experts, and remains difficult to scale as soil degradation increases [European Court of Auditors, 2018]. At the same time, modern machine learning approaches for soil horizon identification depend on reliable human annotations for training and evaluation.

In our prior work we introduced SoilNet, a multimodal multitask model for end-to-end soil horizon prediction from profile images [Singh *et al.*, 2026; Chiaburu *et al.*, 2025]. The modularized pipeline is inherently designed to be transparent by following the soil survey workflow human experts developed for this task. It predicts horizon boundaries followed by morphological horizon features, such as humus content or soil texture, and finally assigns complex hierarchical horizon symbols composed of suffixes and mixtures, using a graph-based taxonomy. However, there is limited research on the difficulty of the task based on image-only inputs, the variability among expert annotators, and the potential for human-AI collaboration in this domain.

To address these gaps, we developed the SoilNet-Annotation App, a web-based tool to collect expert annotations of soil horizons in soil profile images. The app focuses on image-only annotation to mirror the model inputs and standardize the workflow across annotators. It supports boundary placement (segmentation) and hierarchical horizon labeling in a single guided interface, enabling consistent data capture and inter-annotator analysis, see Figure 1. The demonstrator builds on the German Agricultural Soil Inventory (BZE-LW) imagery curated by the Thünen Institute [Thünen Institute, 2026] and provides the experimental backbone for evaluating SoilNet against human experts.

To provide a solid scientific backing for the SoilNet App and build trust in the technology developed we conducted a user study in which our SoilNet App was used by domain experts as assistive technology. These studies allowed for a comparison of human expert performance with and without AI assistance provided by SoilNet. Beyond this user study, the SoilNet-Annotation App offers a workflow and interface design that can be adapted for future annotation efforts in soil science. It serves as a prototype for expert annotators to try out annotating soil profiles together with AI assistance.

2 Soilnet Annotation App

This section provides an overview of the workflow, the user interface design and some implementation details of our publicly-available[1] web-based SoilNet Annotation App.

The workflow is illustrated in Figure 1 and can be broadly structured in two sub-tasks: Segmentation into soil horizon layers and annotation of soil horizon symbols. After a short onboarding screen with a screencast that serves as a tutorial for new users, the users can select whether they want to start the annotation process with or without AI assistance. In the AI-assisted mode, SoilNet-generated horizon boundary suggestions and labels are loaded for user review and adjustment.

[1]https://soilnet.demo.calgo-lab.de

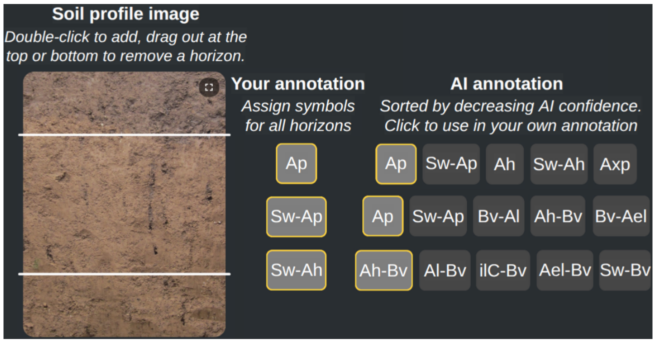

Figure 1: Annotation Interface of the Soilnet-Annotation App[1]

Task 1: Segmentation Regardless of the mode, users proceed to annotate the profile image, by creating horizon boundaries by double-clicking on the image and then refine them by dragging. Boundaries can be removed by dragging them out of the image area.

Task 2: Soil Horizon Annotation After segmentation, users assign horizon symbols via a hierarchical label menu that constrains choices to valid taxonomy paths. Alternatively in AI-assisted mode, users can select one of the Top-5 predictions from our SoilNet model. During this mode, users can update the horizon labels predicted by the model after adjusting the boundaries. SoilNet will use the updated boundaries as input to re-predict the horizon labels. Once satisfied, users can submit their annotations and proceed to the next profile or end the survey session. Users can also skip profile images.

After ending the survey session, users can give feedback on their experience with soil horizon annotation and the app itself. Finally, users can compare their annotations with the ground truth derived from field analysis.

Interface Design The annotation interface (Figure 1) features a central canvas displaying the soil profile image with boundary lines overlaid. A side panel provides hierarchical symbol selection, while primary actions (submit, skip, end) remain persistent and clearly separated. The SoilNet predictions are displayed clearly separated to the right of the user's symbol selection. For precise placement, the interface offers a full-screen zoom view and an overview panel to keep global context. The UI text is presented in German to match the language of participating surveyors, while the demo presented here uses an English localization.

Implementation Details The demonstrator consists of a frontend web application built with React and a backend server implemented in Python using FastAPI. The frontend handles user interactions, while the backend manages data storage, image sampling, and the execution of SoilNet for AI-assisted annotations. An SQLite database stores user annotations and feedback, while the anonymized user ID is saved in the session cookies. The frontend is exported as static files, which are then served by the backend server. The application is containerized using Docker and deployed on our Kubernetes cluster at the Berlin University of Applied Sciences. For resource efficiency, the container running in the backend server only uses CPU resources, as the inference of SoilNet on single images is lightweight enough to not require GPU accelaration. For details on the SoilNet model architecture and training, we refer to our main paper [Singh *et al.*, 2026].

3 User Study

This section summarizes the study design and introduces the evaluation metrics used to quantify performance and agreement. To ensure data integrity, the study was conducted on an instance of the app, separate from the public demo version.

Experimental Set Up We conducted a user study with experienced soil surveyors, comparing manual annotation against AI-assisted annotation using SoilNet. Table 1 summarizes the participation statistics for both phases of the study.

	Phase I	Phase II	Total
Participants	22	23	40*
Profile Images	63	104	142*
Annotations	172	240	412
True Horizon Labels	682	988	1670
Annotator Horizon Labels	695	1053	1748

*Total represents unique counts; some participants and images overlapped between phases.

Table 1: Overview of the User Study Statistics. The study was conducted in two phases: Phase I (Manual) and Phase II (AI-Assisted).

Profile Image Sampling To ensure a fair comparison between human experts and SoilNet, profile images are sampled from the same test split used in our model evaluation [Singh *et al.*, 2026], so that the model has not seen these images during training. The test split contains 670 profile images with roughly the same distribution of horizon labels as the full dataset. Images are randomly drawn from this pool, while prioritizing profile images with existing annotations. Once a profile image reached three independent annotations, it is removed from the sampling pool. We also ensure that no annotator is shown the same profile image twice, by assigning anonymized annotator IDs. This approach allows us to quantify inter-annotator agreement, by comparing multiple independent annotations on the same profile image.

Evaluation Metrics To evaluate the performance of users and the model against field annotations, used as true data we use *1D-Intersection over Union (IoU)* scores, the *accuracy* of the horizon symbol and the *hierarchical main symbol accuracy* that reflects the soil horizon taxonomy.

To quantify annotation consistency, we computed inter-rater agreement metrics for all soil horizons where at least three independent experts provided annotations ($n \geq 3$). We adopt the percent agreement, well-known for simplicity and interpretability [McHugh, 2012]. The metric measures the consensus among our users, calculated as the number of agreements in responses, divided by the sum of agreements and disagreements in the responses. In the following we refer to this as the *Inter-User Agreement (IUA)*, expressed as a percentage from 0% (total disagreement) to 100% (total agreement). To assess the utility of the AI suggestions for experts, we further evaluated the *User-Model Agreement (UMA)*. This quantifies the alignment of the model's predictions with expert decisions. We calculated this as the percentage of expert annotations that matched the model's output. We report this alignment for both the model's Top-1 prediction and its Top-5 prediction set. The reported metrics are the averages across all horizons, for both full and hierarchical main symbols.

4 Results

Figure 2 compares assisted and non-assisted users against the model. The results are weighted by the number of annotations each user contributed. By doing that all profile images annotated by the users have an equal impact on the final results. The error bars show the standard error of the mean (SEM). The evaluation of IUA and UMA are summarized in Table 2.

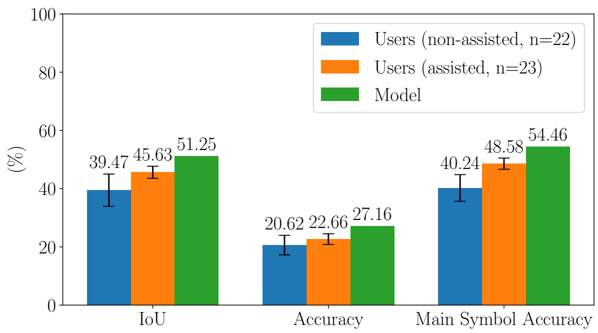

Figure 2: User vs. Model Performance. The model consistently achieves the highest scores, while AI assistance improves user performance across all metrics. To ensure each image has an equal impact, results are weighted by the number of annotations per user. Error bars represent the standard error of the mean (SEM).

Metric	Hierarchical Level	AI-Assisted?	
		No	Yes
IUA	Full	36.86	39.94
	Main	58.43	67.10
UMA	Full	–	15.45
	Main	–	52.12
Top-5 UMA	Full	–	44.25
	Main	–	60.99

Table 2: Inter-User and User-Model Agreement. The table summarizes the average percent agreement among annotators and between annotators and the model, highlighting alignment with model predictions and the impact of model assistance on consistency.

5 Discussion

We now discuss what the results imply for image-only horizon annotation and human-AI collaboration, and we highlight limitations that affect the interpretation of the study.

Performance and Collaboration Effects The results presented in Figure 2 indicate that soil horizon annotation from image-only inputs is a challenging task, even for experienced experts. The model consistently outperforms both assisted and non-assisted users across all metrics. However, AI assistance enhances user performance, narrowing the gap with the model. This suggests that SoilNet's predictions provide valuable guidance to experts, improving their annotation accuracy and consistency. The findings highlight the potential of human-AI collaboration in soil horizon annotation, where AI can augment expert capabilities and lead to more reliable data collection. Note that we simplified the label taxonomy and mapped the original 1218 horizon symbols to 99 horizon symbols. This reduction was necessary to make the learning feasible to the model, as many horizon symbols had only very few examples. These simplifications were then double-checked by our horizination experts.

Agreement Analysis The IUA improves with AI assistance, indicating that SoilNet's predictions help standardize annotations among experts. The UMA metrics further delineate that assisted users align similarly with the model's predictions, when considering the Top-5 predicted labels. This suggests that AI assistance not only enhances individual performance but also promotes consensus among annotators.

Limitations First, the two-phase design introduces temporal effects and may cause alterations in the participant pool. Since user identification relied on browser cookies - which are subject to deletion - we could not reliably track individuals across phases. Over the two phases we can only confirm 5 users that participated in both phases, meaning that the performance gap could stem from the difference in user experience. Second, although images were drawn from the same test split, the specific image subsets differed between phases due to sampling variance. Next, the simplified taxonomy (reduced to 99 classes) differs from the full pedological standard, potentially forcing experts to adapt their natural labeling granularity to the model's constraints. Finally, we did not analyze potential automation biases, where experts blindly trust the AI suggestions, which is part of future work.

6 Conclusion and Future Work

The SoilNet-Annotation App makes the challenges of image-only horizon annotation noticeable while providing an interface for structured data collection. Our user study shows that while this task remains difficult for experts without in-field context, human-AI collaboration improves both performance and inter-annotator agreement. AI assistance not only boosted accuracy metrics but also fostered consensus on hierarchical main symbols, increasing the alignment between human experts and the model's predictions. These findings suggest that deploying models like SoilNet as interactive assistants can standardize large-scale soil monitoring efforts. Future work will explore integrating explanatory features into the app to further support expert decision-making and refining the label taxonomy to better bridge the gap between machine learning constraints and pedological practice. Finally, we will also analyze potential automation biases to ensure experts maintain critical independence from AI suggestions.

Acknowledgments

The authors gratefully acknowledge the financial support provided by the the German Research Foundation (DFG) - Project number: 528483508 - FIP 12.

References

[Ad-hoc-AG Boden, 2005] Ad-hoc-AG Boden. *Bodenkundliche Kartieranleitung, 5. Auflage [KA5 - Soil Survey Guidelines, 5th ed.]*. Schweizerbart Science Publishers, Hannover, Germany, 2005.

[Chiaburu *et al.*, 2025] Teodor Chiaburu, Vipin Singh, Frank Haußer, and Felix Bießmann. Uncertainty-guided expert-ai collaboration for efficient soil horizon annotation. *ECAI 2025, CLEAR-AI Workshop, Bologna*, 2025.

[European Court of Auditors, 2018] European Court of Auditors. Background paper: Desertification in the eu, 2018. Accessed: 1.02.2025.

[FAO, 2015] FAO. *Status of the World's Soil Resources: Main Report*. Food and Agriculture Organization of the United Nations, Rome, Italy, 2015.

[Hartemink *et al.*, 2020] A.E. Hartemink, Y. Zhang, J.G. Bockheim, N. Curi, S.H.G. Silva, J. Grauer-Gray, D.J. Lowe, and P. Krasilnikov. Chapter three - soil horizon variation: A review. volume 160 of *Advances in Agronomy*, pages 125–185. Academic Press, 2020.

[IUSS Working Group WRB, 2022] IUSS Working Group WRB. *World Reference Base for Soil Resources: International soil classification system for naming soils and creating legends for soil maps*. International Union of Soil Sciences (IUSS), Vienna, Austria, 4th edition, 2022.

[McHugh, 2012] Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012.

[Singh *et al.*, 2026] Vipin Singh, Teodor Chiaburu, Einar Eberhardt, Stefan Broda, Joey Prüssing, Frank Haußer, and Felix Bießmann. SoilNet: A multimodal multitask model for hierarchical classification of soil horizons. *Geoderma*, 466:117684, February 2026.

[Thünen Institute, 2026] Johann Heinrich von Thünen Institute: Federal Research Institute for Rural Areas, Forestry and Fisheries Thünen Institute. German Agricultural Soil Inventory (BZE-LW), 2026. Accessed: 2026-01-30.