

The MESOSCAN dataset: a collection of mesofauna images from UK forest and peatland soils

Timothy C. Lukins ^{*}, Thomas Ashton , Ainoa Pravia 

Northern Research Station, Forest Research, Roslin, EH25 9SY Midlothian, United Kingdom

^{*}Corresponding author. Forest Research, Northern Research Station, Bush Estate, Roslin, Midlothian, EH25 9SY, UK. E-mail: tim.lukins@forestresearch.gov.uk

Abstract

Mesofauna such as *Acari* and *Collembola* represent a key bioindicator of soil health and condition, but limited availability of taxonomic expertise and the time-intensive nature of their identification constrain our ability to characterize their community structure and function in forest and peatland ecosystems. As these limitations stem largely from challenges in species identification, novel machine learning (ML) approaches offer promising tools to address them, but in turn require substantive collections of annotated images. Recent large-scale datasets have limited mesofauna images, and to address this paucity in training data we have developed the complementary MESOSCAN dataset with samples taken from UK forest and peatland soils, using an optimized macro-photography system to capture high-resolution and multi-orientation images of specimens to provide broad representation of the main UK mesofauna groups. We validate this resulting dataset by showing how it can be used to construct a simple baseline ML model to confirm good predictive accuracy. The dataset can be found at: <https://doi.org/10.5281/zenodo.15755210>.

Keywords soil, mesofauna, machine learning, forest, peatland

Introduction

Soil mesofauna

Soils are one of the most diverse habitats in the world, estimated to host 59% of the species on Earth (Anthony *et al.*, 2023). Soil fauna drives key essential processes, such as decomposition, mineralization, and nutrient cycling (Bardgett *et al.*, 2005, Nielsen *et al.*, 2011), alongside mediating important ecosystem services such as water purification and retention, climate regulation, pest, and disease suppression (Dominati *et al.*, 2010).

The spatio-temporal distribution and diversity of soil fauna are constrained by the highly complex heterogeneous nature of soil habitats, which are composed of mineral particles, air- and water-filled pores, decomposing and living organic matter, and diverse soil biota (Nielsen, 2019). Further constraints on the diversity of soil fauna communities are imposed by aboveground interactions, including vegetation and terrestrial fauna, which influence the abiotic conditions of soil habitats, such as microclimate and physicochemical properties (Wu *et al.*, 2011). This complex environment is home to the wide range of soil organisms—from microfauna (20 μ m–200 μ m) to macrofauna (2 mm–20 mm) and megafauna (> 20 mm).

Mesofauna (200 μ m–2 mm) exist in-between and are amongst the most studied soil taxa, yet substantial gaps remain in our understanding of their diversity, ecology, and functional roles (Guerra *et al.*, 2020, Thakur *et al.*, 2020). Although standardized and relatively

straightforward sampling and extraction methods exist (Potapov *et al.*, 2020), their small body size coupled with the limited availability of comprehensive identification resources (e.g. keys only to family level) continues to hinder their study at finer taxonomic scales, such as species level.

The bioindicator value of mesofauna

Collembola (Hopkin, 1997) and *Acari* (Walter and Proctor, 2013) are among the most abundant and diverse components of soil mesofauna and are widely recognized as sensitive indicators of environmental change. Both groups generally exhibit higher abundance and diversity in soils with high organic matter content, although community responses are strongly influenced by local factors including soil moisture, habitat structure, and disturbance regime (Querner and Bruckner, 2010, Joimel *et al.*, 2021). Peatlands frequently support distinctive assemblages characterized by specialized taxa to acidic, waterlogged, and anoxic conditions. Consequently, the ecological value of peatland mesofauna often lies on the unique nature of the assemblages rather than high diversity.

The ecological response of mesofauna can be detected across multiple levels of taxonomic resolution, providing flexibility for their use as bioindicators. Environmental gradients and land-use effects have been successfully characterized using classifications from order to superfamily level (George *et al.*, 2017), whilst species to genus level

Received: 22 January 2026. **Revised:** 8 July 2026. **Accepted:** 14 August 2026

© The Author(s) 2026. Published by Oxford University Press on behalf of the Institute of Chartered Foresters.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

can identify key drivers of community response to disturbance (Manu *et al.*, 2021), and the overall richness of oribatid mite species reveal relationships with soil organic matter and vegetation (Pravia *et al.*, 2025). These findings suggest that meaningful ecological information can be extracted from different levels of taxonomic identification in mesofauna communities.

Forest management practices can produce significant spatio-temporal changes in mesofauna communities, with measurable impacts on soil quality detectable at the species level (Heydari *et al.*, 2020). Among mesofauna taxa, oribatid mites provide useful insights into the dynamics within forest ecosystems, as their community structure is strongly influenced by seasonality, microhabitat composition (Wehner *et al.*, 2018), and tree diversity through linkages with litter decomposition and organic matter. Owing to their high abundance, taxonomic diversity, and strong association with organic matter in soils, oribatid mites are frequently regarded as reliable indicators of soil quality in both forest and peatland ecosystems (Hågvar *et al.*, 2020). These relationships are however not always detectable within all mesofauna groups, e.g. *Mesostigmata* and *Collembola* (Wenglein *et al.*, 2025).

Coarser taxonomic resolution (family and above) can be used to assess for example recovery following wildfire in boreal forests (Meehan *et al.*, 2019) and to evaluate the effects of forest management interventions such as thinning and prescribed burning (Gibson *et al.*, 2022). In land use conversion systems, including transitions from arable to plantation forestry, species-level diversity metrics of collembolan communities have demonstrated sensitivity to vegetation composition and residual soil fertility, highlighting their potential as bioindicators of ecosystem change (Harta *et al.*, 2021).

Oribatid mites also exhibit strong responses to environmental change in peatlands, showing to be strongly linked to water organic carbon, making them complementary bioindicators of water quality (Seniczak *et al.*, 2022). Peatland drainage associated with forestry, a common historical disturbance across UK and Ireland, has been linked to reductions in mesofauna abundance and shifts in community composition. These shifts have been detected at relatively coarse taxonomic resolutions (order to suborder) and are strongly associated with water table depth and nutrient status (Silvan *et al.*, 2000). The sensitivity of mesofauna communities to changes in peatland hydrology highlights their potential application in peatland restoration monitoring. However, evidence suggests that faunal recovery often lags hydrological recovery in peatland restoration, with community reassembly occurring over longer timescales than changes in hydrology (Fundira *et al.*, 2025). Furthermore, some responses can frequently be context-dependent, reflecting interactions between temperature, vegetation composition, and peatland hydrology (Barreto *et al.*, 2021). This sensitivity to multiple environmental drivers enhances the value of mesofauna as bioindicators capable of reflecting both abiotic conditions and broader ecosystem recovery.

Challenges studying mesofauna communities

The study of mesofauna communities typically relies on standardized sampling and extraction protocols. A commonly used extraction method employed is Tullgren extraction, in which soil samples are placed on a fine mesh beneath a heat source, creating a temperature, and moisture gradient that drives organisms downward into a collection vessel containing a preservative (Southwood, 1978). These methods are relatively straightforward and facilitate the processing of

large numbers of samples, with throughput primarily constrained by the number of extraction funnels available and the duration.

The principal limitation in mesofauna research arises after extraction, particularly specimens sorting and taxonomic identification. Accurate species identification requires both specialist taxonomic expertise and the availability of reliable identification keys. The development of such keys depends on a sufficiently resolved taxonomy, which in turn relies on comprehensive ecological knowledge of the target group, often supported by national species inventories and checklists. Even when suitable taxonomic resources are available, the processing of a single sample can require several days of work by an experienced taxonomist, depending on the abundance and diversity of specimens present.

The labour-intensive and specialist nature of taxonomic identification can limit both sampling replication and the taxonomic resolution achieved within studies. These constraints are further compounded by the inherent spatial heterogeneity of soil habitats, which can obscure ecological patterns and complicate the interpretation of community-level responses and trends. Gaining more spatial information can only be achieved through improvements in sampling throughput, again limited by a scalable solution to identification.

Machine learning datasets and deficiencies

Recent advances in using techniques from artificial intelligence (AI) have proven to be a fertile area of applied research in ecology, and with application to soil science (Wadoux, 2025). A particular example of this is the subfield of AI that is machine learning (ML).

A useful introduction to applied ML for ecologists is presented by Blair *et al.* (2024) and by Høye *et al.* (2021) who highlight the potential of ML to advance biodiversity studies, specifically by collecting sufficient image data along with labelled identification and further incorporating additional contextual information to help improve model performance. These insights informed the recent work of Gharaee *et al.* (2023) in developing the BIOSCAN project which showed what can be accomplished on a large scale using datasets of up to 1 million images (later 5 million), including taxonomic and molecular data. The initial BIOSCAN-1M dataset was consequently combined with a state-of-the-art CLIP visual language model approach (Radford *et al.*, 2021) to build the CLIPD system to show how by incorporating multimodal data it was possible to further improve classification (Gong *et al.*, 2025). Even larger image datasets have also been constructed in a similar fashion and have also been used for classification, e.g. BioCLIP (Stevens *et al.*, 2023) has 10 million images in 450 000 taxa and BioTrove (Yang *et al.*, 2024) which has 161 million images for 366 600 species.

All of these datasets have a principal focus on larger invertebrates, often have limited fine taxonomic resolution (i.e. down to genus level), can include images that may suffer from quality issues, and may also include bias in the representation of taxa covered. For example, BIOSCAN-5M includes 46 544 records for *Collembola* of which only one order is represented (*Entomobryomorpha*, 35 982 records) belonging to one family (*Entomobryidae*, 20916 records). Of the 24 species represented in this family, 4317 or 10% of the total records for *Collembola* correspond to the species *Entomobrya nivalis*, all of which were collected from one site in Sweden. For *Acari*, the order *Sarcoptiformes* contains 5645 records and 50 families that are unassigned to finer taxonomic levels with the exception of *Oribatulidae* (*Phauloppia lucorum* contains 59 records).

Specific machine learning mesofauna image datasets

These large image-based datasets for arthropod identification have achieved broad taxonomic coverage of larger arthropod (and other) taxa but tend to under-represent smaller taxa, including soil mesofauna. This bias likely reflects the predominance of datasets derived from single images of individual specimens, which are often sufficient for species-level identification of larger organisms but may lack the resolution necessary to capture the diagnostic morphological features required for reliable identification of smaller taxa. This is mainly due to the fact that it is extremely challenging to provide sufficiently well curated data for soil mesofauna—in terms of the level of training data, detail of taxonomic identification, and quality of images for such small organisms.

To address these challenges associated with species-level identification of small-bodied taxa, Schneider *et al.* (2022) developed a computer vision imaging approach based on broad arthropod functional groups for the estimation of abundance, biomass, and diversity. Their dataset comprised 13 059 cropped specimen images representing 18 arthropod groups and 2 debris classes, including 312 images of *Acari* and 418 images of *Collembola*. Classification performance for these mesofauna groups was lower with accuracies of 80% compared with 86%–96% achieved for larger arthropod groups. The reduced performance was attributed to a combination of small body size, limited image representation of training data, and substantial size heterogeneity within *Acari* and *Collembola*, forcing the classifier to recognize highly variable morphologies into a single category.

In response to this, the seminal system—which inspires this work—CollembolAI (Sys *et al.*, 2022) showed how specific and better annotated training data could lead to successful species-level classification of mesofauna specimens using curated datasets. Sys *et al.* used a digital SLR camera and a macro-lens, mounted on a CNC/Cartesian robot arm (as often found on 3D printers) to capture images, primarily of ten species of *Collembola* alongside two species of *Acari*. Image capture involved a succession of overlapping images across 14 cm Petri dish samples, resolving to $100\ \mu\text{m} \approx 36\ \text{pixels}$ to create 6113 individually annotated images for training.

Oriol *et al.* (2024) compiled a composite *Collembola* dataset from multiple agricultural studies, employing higher magnification microscopy (resolving to $100\ \mu\text{m} \approx 300\ \text{pixels}$) to build a dataset of 2195 annotated images. Similarly Zhuang *et al.* (2025) also focused on *Collembola* identification using deep learning methods based on a dataset of 5020 individuals that span 6 families, 10 genera, and 51 species. They were imaged with a stereo microscope at 50 $\times$ magnification with specimens used in a variety of positions with a minimum of 23 images per species.

In addition to these approaches, it is also worth noting the development of other automated high-throughput solutions such as the Edapholog system (Flórián *et al.*, 2020) which represent more specialized approaches to achieve high-speed detection and classification. The resulting lack of high-resolution data means that only higher-order taxa are distinguishable, but at the trade-off of greater speed. They also illustrate the challenges associated with development of solutions using custom hardware, which is often reflected in the resulting cost of such a system.

More recently, the development of the COLLEMBOT system (Wehrli *et al.*, 2026) focused principally on performing counts of adult and juvenile springtails of the species *Folsomia candida* in accordance with standardized ecotoxicological assessments. This was developed

with both a Mask R-CNN and YOLOv11 convolutional neural network models trained on 3207 images selected from five different existing datasets (including CollembolAI). Their results contrast the two models revealing systematic differences in detection performance that depended on image quality and effects of scaling as critical determinants of the model's counting and segmentation performance.

As a useful reference to the current state-of-the-art, Remelli *et al.* (2026) recently performed a synthesis review focuses on how traditional taxonomy, molecular, and ML approaches can be combined to overcome the taxonomic impediments that limit large-scale soil biodiversity monitoring—highlighting their strengths and weaknesses. The main conclusion being that they should not be seen as competitive but rather as complementary, that instead integrating morphological expertise, genetic information, and ML, represented the most promising approach to scalable, efficient, and reliable assessment of soil fauna.

Aims of this dataset

The main contribution of our MESOSCAN dataset presented here is to build on these previous efforts to provide additional example images to help identify and automate the taxonomic identification of mesofauna. It uses specimens from a wide range of habitats, primarily forests, and peatlands to create a morphologically representative collection of UK mesofauna. This image collection is not automated nor high-throughput, but presents a useful methodology to enable the compilation of suitable good quality images, and to grow this to include further specimens. By implementing and presenting a reference benchmark ML system alongside the dataset we seek to demonstrate that the data itself contains suitable discriminatory information and is fundamentally fit-for-purpose.

Materials and methods

Sample collection

Samples are recorded against an original `SAMPLE_ID` (TrueType font indicates a metadata field recorded as a column in the accompanying dataset `MESOSCAN.csv` file) and were collected from 21 sites (`SAMPLE_SITE`) across the UK. The sites covered a variety of habitats (`SAMPLE_HABITAT`) corresponding to UKHab coding, and were noted for the soil type (`SAMPLE_SOIL`) as defined according to the National Map of Scotland (Soil Survey of Scotland Staff, 1981) and the Forestry Commission Soil Map of England (Pyatt, 1982).

The source of samples (`SAMPLE_PROJECT`) is indicated, with the majority of forest associated specimens taken as part of the UK Government Natural Capital and Ecosystem Assessment (NCEA) programme and with peatland examples taken from a range of sites as part of other work.

Forest samples were collected between 2023 and 2025 during autumn and winter field seasons, with peatland collections performed at variable times of the year (`SAMPLE_DATE`). Location is provided as coordinates (`SAMPLE_LAT` & `SAMPLE_LON`) (Fig. 1a). The exact locations of sample surveys conducted under the NCEA programme are confidential—and are de-located as the centroid of a bounding box where each bounding box contains a minimum of 200 hectares of woodland, with the survey taking place at an undefined location within this area.

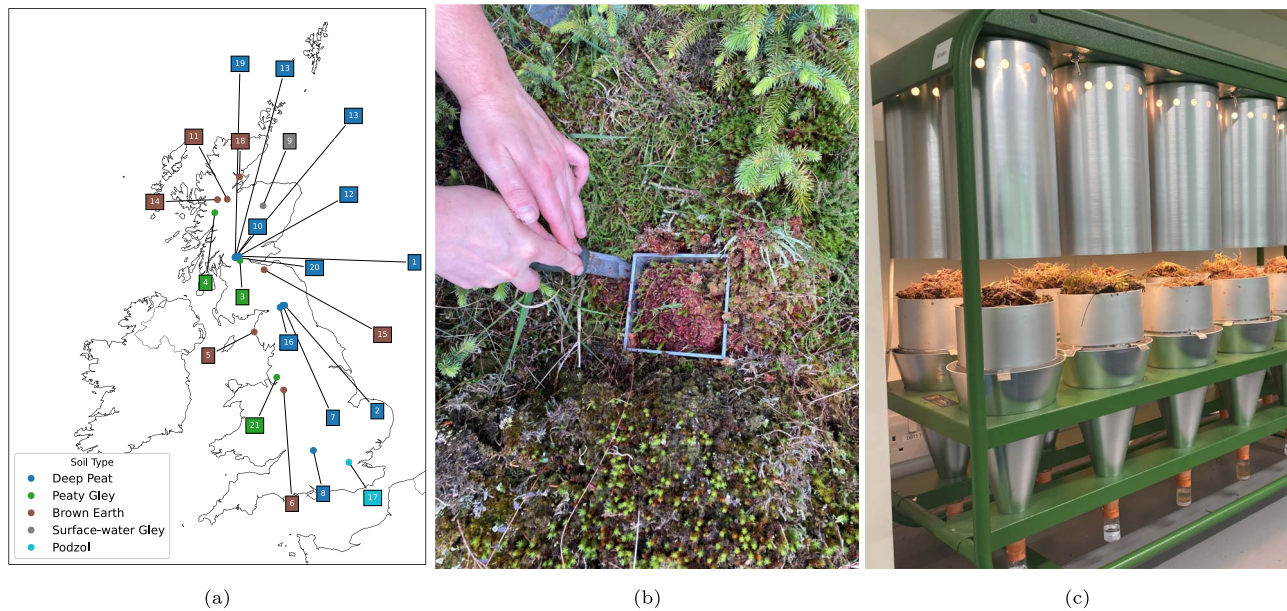


Figure 1 (a) Map of the UK showing location of 21 sample sites by soil type; (b) collection of peatland samples using a quadrat to extract 10 cm^3 ; (c) mesofauna extraction using Tullgren funnels where specimens are collected into the lower vials preserved in 96% molecular graded ethanol for later identification.

Soil samples in peatlands consisted of a $10 \times 10 \times 10\text{ cm}$ quadrat sample (Fig. 1b), whilst forest samples consisted of an $8 \times 10\text{ cm}$ diameter soil core taken with an AMS corer (AMS Inc. Idaho, USA. <https://www.ams-samplers.com/soil-core-samplers/>). The quadrat approach for peatlands is necessary due to the very wet soil conditions, where an impact corer is ineffectual, and so samples are collected swiftly as possible to ensure highly mobile, surface-active specimens are captured. As this dataset aims to capture the morphological diversity of UK mesofauna, differences in sampling methodology between habitat types does not affect the final dataset, as specimens have been selected to represent morphological diversity. All specimens were extracted in Tullgren funnels (Fig. 1c).

Specimen selection and identification

Unlike other mesofauna datasets, our approach images single species instead of whole samples, and thus specimens were identified first to the finest taxonomic resolution possible. Taxonomic identification relied on published (Shepherd and Crotty, 2024) and unpublished *Acari* keys, and established *Collembola* keys (Hopkin, 2007). Six to ten specimens of the same taxonomic rank were then selected for individual image capture. The exact number of individuals used from each sample is listed (NUM_INDIVIDUALS). Crucially, each imaged specimen had been taxonomically identified beforehand, bypassing the need of annotation which can involve imaging the entire sample and then labelling and defining bounding boxes around individuals.

When considering the coverage of UK mesofauna, specimen selection for the MESOSCAN dataset is guided by three principles:

1. Availability of enough specimens of the same species in the same sample.
2. Coverage across all higher taxonomic groups.
3. Representing morphological diversity within *Acari* and *Collembola*.

To further guide selection, and to harmonize the classification, taxonomic ranks were ordered by TAXON_CLASS (*Collembola* and *Acari*), and TAXA_GROUP levels of taxonomic classification: for *Acari*, the Orders *Mesostigmata*, Suborders *Oribatida*, *Prostigmata*, *Endeostigmata*, and Cohort *Astigmata* (Shepherd and Crotty, 2024); for *Collembola*, the Super-families *Entomobryomorpha*, *Poduromorpha*, *Symphyleona*, and *Neelipleona* (Hopkin, 2007).

TAXA_GROUP is followed by finer taxonomic groups (TAXA_FAMILY, TAXA_GENUS, and TAXA_SPECIES). A unique code is allocated to each taxon class (TAXON_ID).

The structure of the data aims to cover morphological diversity across UK mesofauna groups, and provide a classification structure that helps normalize and assess bias in the data, assist curation and be expanded with additional ranks if required (Table 1). The predominance of *Oribatida* within the dataset reflects both their exceptional taxonomic diversity (60–65 families in the UK, plus 230–240 estimated families worldwide) and their ecological importance in forests and peatlands. The dataset also includes immature life stages, nymphs, which can be abundant in sampled communities, particularly for *Oribatida*. Often, the diagnostic morphological characters required for reliable species-level identification are absent during early nymphal stages, but nymphs account for a substantial fraction of total oribatid mites found in sampled communities and so ideally need to be accounted for.

Image capture

The imaging capture system (Fig. 2) consists of a Canon EOS-90D digital SLR camera equipped with a Canon MP-E 65mm f/2.8 1-5 $\times$ lens, and a Canon Speedlite MR-14EX II macro ring flash (Canon Inc., Tokyo, Japan. <https://global.canon/>). The camera is mounted on a WorkBee Z1+ CNC machine (Ooznest Ltd., Brentwood, UK. <https://ooznest.co.uk/>) using a simple L-bracket with standard 3/8–16 tripod screw size to attach the camera. An additional dual swan neck microscope

Table 1 Dataset taxa selected for image capture, organized by simplified taxonomic structure split principally by *Collembola* and *Acari* with varying resolution of identification and the addition of nymphs as indicated by (N).

Class	Group	Family	Genus	Species
<i>Acari</i>	Mesostigmata	Trachytidae	<i>Trachytes</i>	<i>aegrota</i>
<i>Acari</i>	Mesostigmata	Uropodidae	<i>Uropoda</i>	<i>minima</i>
<i>Acari</i>	Mesostigmata	Rhodacaridae	<i>Rhodacarellus</i>	<i>silesiacus</i>
<i>Acari</i>	Mesostigmata	Parasitidae	<i>Paragamasus</i> sp.	
<i>Acari</i>	Mesostigmata	Veigaiidae	<i>Veigaia</i>	<i>nemorensis</i>
<i>Acari</i>	Astigmata agg.(N)			
<i>Acari</i>	Oribatida	Galumnidae	<i>Acrogalumna</i>	<i>longipluma</i>
<i>Acari</i>	Oribatida	Scheloribatidae	<i>Liebstadia</i>	<i>similis</i>
<i>Acari</i>	Oribatida	Phthiracaridae	<i>Steganacarus</i>	<i>michaeli</i>
<i>Acari</i>	Oribatida	Phthiracaridae	<i>Phthiracarus</i> sp.	
<i>Acari</i>	Oribatida	Brachychthoniidae	<i>Sellnickochthonius</i>	<i>zelawaiensis</i>
<i>Acari</i>	Oribatida	Nanhermanniidae	<i>Nanhermannia</i>	<i>coronata</i>
<i>Acari</i>	Oribatida	Nanhermanniidae	<i>Nanhermannia</i>	<i>coronata</i> (N)
<i>Acari</i>	Oribatida	Malaconothridae	<i>Malaconothrus</i>	<i>monodactylus</i>
<i>Acari</i>	Oribatida	Euzetidae	<i>Euzetes</i>	<i>globulus</i>
<i>Acari</i>	Oribatida	Hypochthoniidae	<i>Hypochthonius</i>	<i>rufulus</i>
<i>Acari</i>	Oribatida	Euphthiracaridae	<i>Acrotritia</i>	<i>duplicata</i>
<i>Acari</i>	Oribatida	Opipiidae	<i>Opipella</i>	<i>nova</i>
<i>Acari</i>	Oribatida	Quadropiidae	<i>Quadroppia</i> sp.	
<i>Acari</i>	Oribatida	Camisiidae	<i>Platynothrus</i>	<i>peltifer</i>
<i>Acari</i>	Oribatida	Camisiidae	<i>Platynothrus</i>	<i>peltifer</i> (N)
<i>Acari</i>	Oribatida	Achipteriidae	<i>Parachipteria</i>	<i>punctata</i>
<i>Acari</i>	Oribatida	Peloppiidae	<i>Ceratoppia</i>	<i>bipilis</i>
<i>Acari</i>	Oribatida	Ceratozetidae	<i>Ceratozetes</i>	<i>gracilis</i>
<i>Acari</i>	Oribatida	Nothridae	<i>Nothrus</i>	<i>pratensis</i>
<i>Acari</i>	Oribatida	Nothridae	<i>Nothrus</i>	<i>pratensis</i> (N)
<i>Acari</i>	Endeostigmata	Nanorchestidae	<i>Nanorchestes</i> sp.	
<i>Acari</i>	Prostigmata	Eupodidae agg.		
<i>Collembola</i>	Entomobryomorpha	Isotomidae	<i>Folsomia</i>	<i>penicula</i>
<i>Collembola</i>	Entomobryomorpha	Isotomidae	<i>Isotomiella</i>	<i>minor</i>
<i>Collembola</i>	Entomobryomorpha	Isotomidae	<i>Parisotoma</i>	<i>notabilis</i>
<i>Collembola</i>	Entomobryomorpha	Entomobryidae	<i>Lepidocyrtus</i>	<i>lignorum</i>
<i>Collembola</i>	Poduromorpha	Neaturidae	<i>Friesea</i>	<i>truncata</i>
<i>Collembola</i>	Neelipleona	Neelidae	<i>Megalothorax</i>	<i>minimus</i>
<i>Collembola</i>	Symphyleona	Sminthuridae	<i>Sminthurides</i>	<i>malmgreni</i>

light (Brunel Duel Flexlite, Brunel Microscopes Ltd., Chippenham, UK. <http://www.brunelmicroscopes.co.uk/>) is used to provide a low-level consistent light source to help visually position and move the camera over the specimens. A white plastic A4 sheet is positioned on the CNC spoil-board to provide a light and contrasting background. Black and darker backgrounds were discovered to highlight debris and scratches in the specimen plates—while white provides good backscattering, but less contrast with lighter specimens.

The entire camera and CNC arm system is contained within a panelled enclosure to exclude external lighting effects. Covered ventilation is provided to facilitate the dissipation of ethanol vapour and prevent its accumulation within the enclosure. An adjacent Dell Precision 3660 computer—with an Intel i9-12900K processor and 128 GB of RAM (Dell Technologies Inc., Texas, USA. <https://dell.com>) is used to operate the system.

The control software is a Python application developed to simplify communication directly with the camera using the Canon camera SDK (<https://www.canon.co.uk/business/imaging-solutions/sdk/>) and synchronize movement with the CNC controller firmware (<https://www.reprapfirmware.org/>). The control of the stepper motors support movement along X, Y, and Z axes in increments of 0.1 mm, allowing for very fine-grained control along each axis of motion to position the camera over the specimens.

The images for each taxonomic rank were captured using the same settings to ensure consistent data. The lens is set to ~ 4× magnification (leaving ~ 2 cm of working distance in front of the flash looking down on the specimens) and aperture is fixed to *f*5.6 so to balance the proximity of the lens and maximize the depth-of-field. A microscope calibration slide is used to confirm that the spatial calibration of the zoom is capable of resolving 100µm ≈ 125 pixels (Calibration images by date are in the ./calibration folder). The equivalent of the film speed is set to ISO 100 and the exposure time is set to 1/100 s. Image settings are also fixed to landscape orientation and output selected for large uncompressed JPEG images at a resolution of 6960 × 4640 pixels in 24 bit colour.

Dataset curation

For each taxonomic rank, 6–10 specimens were used for imaging (PROCESS_ID) using nematode counting plates (University of Wageningen, Wageningen, NL. <https://shop.wur.nl/nematology>). These contain all the specimens within a single well of 6.3 cm in diameter and ≈ 0.8 cm deep constructed of semi-opaque plastic which aids with good light diffusion in combination with the flash. The grid pattern present is useful for positioning when moving across the plate, but is not directly referenced. The walls of the well are

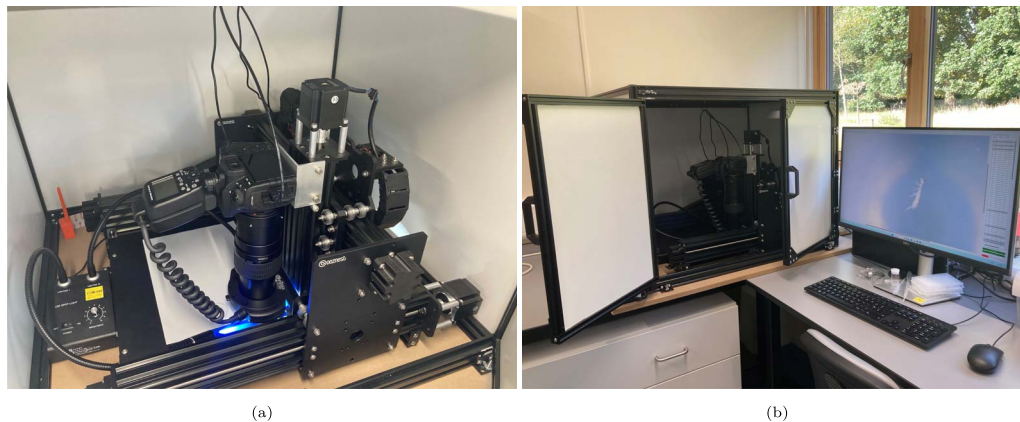


Figure 2 Equipment used, showing internals (a) consisting of digital SLR camera with macro-lens and ring flash mounted on CNC machine arm, alongside modelling light; external view (b) showing enclosure and adjacent control computer used to run control software.

purposefully chamfered to have a shallow angled edge which helps keep the specimens within the centre and not to have the meniscus at the edge obscure and refract light.

All specimens from an taxonomic rank are deposited in the well with ≈ 3 ml of 90% ethanol, and the plate positioned under the camera, by default homed to the bottom left corner origin point. The CNC arm is then guided by the operator to move across the well and capture a series of images when an individual specimen is encountered. At $4\times$ magnification, the field of view is $\sim 5\text{ mm} \times 3.3\text{ mm}$ (in keeping with the image ratio of 3:2)—meaning that the width of the dish can take up to 12 images, and the height ~ 17 images. This can also include over a range of different heights along the Z axis—slowly descending in increments of 0.1 mm. Between runs over the dish, the operator can gently agitate the plate to reposition specimens into new resting positions not captured yet.

Using this approach, a number of photographs were captured over a variety of orientations and degrees of focal depth. From these candidate images the sharpest and most detailed images are selected and cropped to 1600×1600 pixels with the specimen located as close as possible to the centre of the image. (The image filename records the date and time, the XYZ position, and the crop location in the original image—e.g. 250618_153818_X-22-500_Y-42-500_Z-47-500_334_953.jpg). If the same specimen has different areas in focus during multiple exposures, these images are also included (focus stacking was considered but due to movement of the subjects while suspended and the time required to align and create an image, this was deemed impractical). Example images for each taxon class are shown in Fig. 3.

Following this procedure, up to 100 images can be generated for a single taxon in $\sim 3\text{--}4$ hours. All cropped images are then curated into the dataset following the established hierarchical structure shown in Table 1. Totally 2500 images occupy ~ 200 MB of storage (the original images are not retained, this total is for only the cropped images). The specimens are then kept in ethanol vials for further capture if necessary.

Baseline validation

The suitability of the images in the dataset for developing an ML system can be practically demonstrated by constructing a baseline implementation. This is a useful approach to confirm the data has enough discriminatory information present to perform classification,

but is not intended to validate any further analysis or suitability of the model for direct application to the task of identification or wider ecological interpretation. How a model is constructed, further refined, versioned, and deployed into an operational context is mediated by other fine-tuning techniques and approaches.

To demonstrate this, a transfer learning approach based on the MobileNetV2 CNN architecture (Sandler *et al.*, 2018) can be developed using the TensorFlow Keras library (Chollet, 2015) with frozen weights pretrained on the ImageNet dataset of 14 million images. To this can be added a series of data augmentation layers for horizontal or vertical flips and some degree of rotation, a dropout layout with a probability of 0.2 to prevent over-fitting, alongside a final prediction layer with a softmax activation function to map to our taxonomic ranks.

The dataset can then be divided repeatedly using k-fold cross-validation to address any bias and ensure each image is used at least once for training and validation. Stratification, to counteract any class-imbalance, is not required as the distribution of the training data (i.e. number of images) is broadly equivalent per class. We used five-folds resulting in a different split each time of 80% training, 20% validation. We then trained for 40 epochs using an Adam optimizer with a learning rate of 0.0001. As the focus here is proving that a simple model can be constructed, we did not retain an explicit test dataset—instead considering only the validation set. (All code for the production of the model, the analysis, and plots shown here are from the Jupyter notebooks contained within the ./notebooks folder.)

We tracked the accuracy (Fig. 4a)—the proportion of true positive predictions over all cases—and the loss (Fig. 4b)—Sparse Categorical Cross-entropy—during model fitting at each step over each fold of the dataset. Resulting plots indicate the regularity in learning and a final mean across all folds average accuracy of 0.95 (with precision 0.97, recall 0.96, and F1-score 0.97). The trained model performed well, as indicated by classification of the validation set, visualized as a confusion matrix with a strong diagonal response that indicates the majority of specimens are correctly identified.

Discussion

The volume and quality of data are critical to the success of any ML approach. Compared with similar datasets (Sys *et al.*, 2022, Oriol *et al.*, 2024, Zhuang *et al.*, 2025), our objective has been to develop a broad training dataset to support the advancement of ML classification



Figure 3 Example mesofauna images in the dataset for each of the 35 taxon class labels, including commonly encountered nymphs.

systems, with a focus on individual specimens. To evaluate this, we implemented a baseline ML model to demonstrate the feasibility of building a classifier. This serves primarily to validate the discriminatory power of the images as a form of quality control, rather than as a prescriptive solution or a tool for ecological assessment.

In constructing this dataset, we aimed to capture sufficient morphological diversity to enable the model to generalize and identify key features. However, the taxonomic level at which novel images can be

reliably classified remains to be determined. We selected specimens across a wide range of groups, ensuring representation of diverse forms and types. At the same time, we maintained consistent image scale and exposure to maximize comparability across the dataset.

Enhancing the dataset further with additional images of existing species is an obvious step which would lead to improvements through more examples of differing morphology (e.g. size and colour) and orientations.

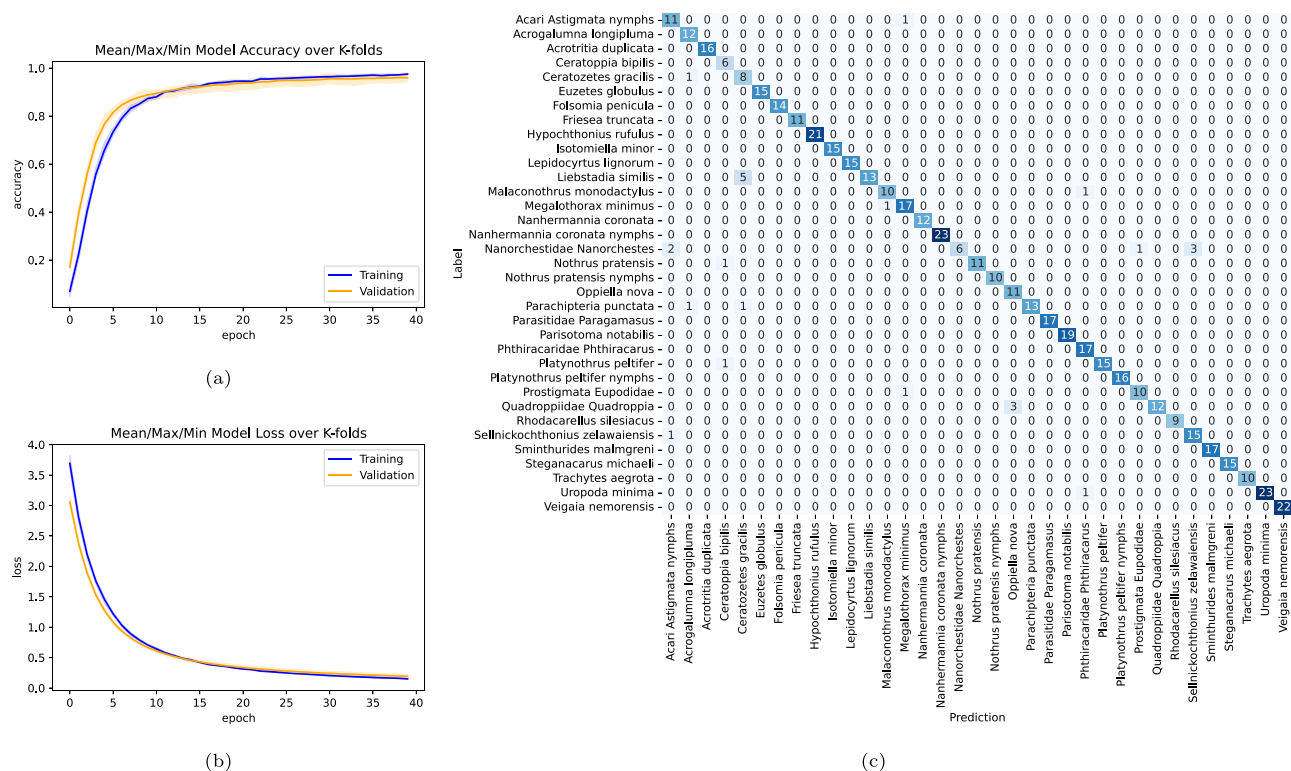


Figure 4 Accuracy (a) and loss (b) during learning over all k-fold ($k = 5$) training runs for both training and validation datasets; alongside confusion matrix plot showing each of 35 class labels and predictions for all 512 instances in random validation set classified by a trained model.

While we have also aimed to ensure a consistent volume of training data by levelling approximately the same number of images per class, further addition of what new species, genus, or family could be incorporated requires careful consideration regarding selection bias for specimen selection, especially in light of the rarefaction and long-tail of actual species encountered (Fabbri et al., 2022). Ultimately, the speed of capturing new samples could be further automated to help scale up data collection. It may also help to consider an alternative colour of background to highlight some of the more lighter colour specimens (in particular *Collembola*).

MESOSCAN is only based on the current morphological and taxonomic information known of mesofauna communities and is, as such, a unimodal dataset. But these communities, as with biodiversity in general, is a complex multidimensional component of ecosystems. In ecological studies it is now common to use functional and molecular information alongside taxonomic and recent work has shown these approaches to be a fruitful direction for significant improvement in automated recognition as it lends itself to current approaches in ML with using multi-modal models that can incorporate this additional information to provide context alongside the images (Gu et al., 2025).

Conclusion

Acari and *Collembola* are important bioindicators of soil quality and health, underpinning critical soil processes, such as decomposition, nutrient cycling, soil structure, and habitat resilience. Understanding mesofauna community structure and composition is important for decoding and maintaining ecosystem functioning, particularly in soil assemblages where they become dominant groups such as

forests and peatlands. Despite their ecological importance, progress in the study of mesofaunal communities remains constrained by the taxonomic impediment, which limits both the speed, robustness, and scalability of soil biodiversity assessments. The MESOSCAN dataset has been developed as a practical resource to support taxonomic expertise whilst facilitating the adoption of ML approaches capable of increasing the efficiency and throughput of mesofauna identification.

Acknowledgements

We thank Conor Strong and Miguel Ibanez for internal reviews of an earlier draft of this paper.

Author contributions

Timothy Lukins (Conceptualization [lead], Data curation, Software, Investigation, Methodology [lead], Validation, Writing—original draft preparation [lead], Writing—review & editing [lead]), Thomas Ashton (Investigation [lead], Data curation [lead]), and Ainoa Pravia (Investigation, Data curation, Funding acquisition, Supervision, Writing—review & editing)

Conflicts of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding

This work was supported by the UK government through the Department for Environment, Food and Rural Affairs (Defra) Natural Capital and Ecosystem Assessment (NCEA) programme for sites in England; and the Department for Energy, Security and Net Zero (DESNZ) for sites in Scotland. Additionally, some work was supported by funding made available from the Scottish, Welsh, and UK government via the Science and Innovation Strategy for Forestry in Great Britain (2021–2025), the Nature for Climate Fund (NCF) Programme, and Forest Research and Development (FRD) Programme.

Data availability

The dataset described by this paper is at: <https://doi.org/10.5281/zenodo.15755210>.

References

- Anthony MA, Bender SF, Van Der Heijden MGA. Enumerating soil biodiversity. *Proc Natl Acad Sci* 2023;**120**:e2304663120. <https://doi.org/10.1073/pnas.2304663120>
- Bardgett RD, Hopkins DW, Usher MB. *et al.* (eds). *Biological Diversity and Function in Soils*. Ecological reviews. Cambridge: Cambridge University Press, 2005. <https://doi.org/10.1017/CBO9780511541926>
- Barreto C, Branfireun BA, McLaughlin JW. *et al.* Responses of oribatid mites to warming in boreal peatlands depend on fen type. *Pedobiologia* 2021;**89**:150772. <https://doi.org/10.1016/j.pedobi.2021.150772>
- Blair JD, Gaynor KM, Palmer MS. *et al.* A gentle introduction to computer vision-based specimen classification in ecological datasets. *J Anim Ecol* 2024;**93**:147–58. <https://doi.org/10.1111/1365-2656.14042>
- Chollet F and others. Keras [computer software]. <https://github.com/fchollet/keras>. 2015.
- Dominati E, Patterson M, Mackay A. A framework for classifying and quantifying the natural capital and ecosystem services of soils. *Ecol Econ* 2010;**69**:1858–68. <https://doi.org/10.1016/j.ecolecon.2010.05.002>
- Fabrizzi S, Papadopoulos S, Ntoutsis E. *et al.* A survey on bias in visual datasets. *Computer Vision and Image Understanding* 2022;**223**:103552. <https://doi.org/10.1016/j.cviu.2022.103552>
- Flórián N, Gránicz L, Gergőcs V. *et al.* Detecting soil microarthropods with a camera-supported trap. *Insects* 2020;**11**:244. <https://doi.org/10.3390/insects11040244>
- Fundira T, Arnscheidt J, Jordan P. Evaluating the effectiveness of Forest-to-bog restoration on carbon sequestration, water chemistry, and biodiversity in Irish and British peatlands. *WIREs Clim Change* 2025;**16**:e70016. <https://doi.org/10.1002/wcc.70016>
- George PB, Keith AM, Creer S. *et al.* Evaluation of mesofauna communities as soil quality indicators in a national-level monitoring programme. *Soil Biol Biochem* 2017;**115**:537–46. <https://doi.org/10.1016/j.soilbio.2017.09.022>
- Gharaee Z, Gong Z, Pellegrino N. *et al.* A step towards worldwide biodiversity assessment: the BIOSCAN-1M insect dataset. *arXiv* 2023. <https://doi.org/10.48550/arXiv.2307.10455>
- Gibson KS, Johnson NC, Laturno C. *et al.* Abundance of mites, but not of collembolans or nematodes, is reduced by restoration of a Pinus ponderosa forest with thinning, mastication, and prescribed fire. *Trees, Forests and People* 2022;**7**:100190. <https://doi.org/10.1016/j.tfp.2022.100190>
- Gong Z, Wang AT, Haurum JB. *et al.* CLIPD: Bridging vision and genomics for biodiversity monitoring at scale. *arXiv* 2025. <https://doi.org/10.48550/arXiv.2405.17537>
- Gu J, Stevens S, Campolongo EG. *et al.* BioCLIP 2: Emergent properties from scaling hierarchical contrastive learning. *arXiv* 2025. <https://doi.org/10.48550/arXiv.2505.23883>
- Guerra CA, Heintz-Buschart A, Sikorski J. *et al.* Blind spots in global soil biodiversity and ecosystem function research. *Nat Commun* 2020;**11**:3870. <https://doi.org/10.1038/s41467-020-17688-2>
- Hågvar S. Ecological spotlights on mites (Acari) in Norwegian conifer forests: a review. In: Haouas D, Hufnagel L, (eds). *Pests Control and Acarology*. IntechOpen, 2020. <https://doi.org/10.5772/intechopen.83478>
- Harta I, Simon B, Vinogradov S. *et al.* Collembola communities and soil conditions in forest plantations established in an intensively managed agricultural area. *J For Res* 2021;**32**:1819–32. <https://doi.org/10.1007/s11676-020-01238-z>
- Heydari M, Eslaminejad P, Kakhki FV. *et al.* Soil quality and mesofauna diversity relationship are modulated by woody species and seasonality in semiarid oak forest. *For Ecol Manage* 2020;**473**:118332. <https://doi.org/10.1016/j.foreco.2020.118332>
- Hopkin SP. *Biology of the Springtails: (Insecta: Collembola)*. Oxford: Oxford University Press, 1997. <https://doi.org/10.1093/oso/9780198540847.001.0001>
- Hopkin S. *A Key to the Collembola (Springtails) of Britain and Ireland*. Occasional Publication - Field Studies Council, FSC, 2007. <https://www.field-studies-council.org/shop/publications/springtails-aidgap/>
- Høyte TT, Årje J, Bjerger K. *et al.* Deep learning and computer vision will transform entomology. *Proc Natl Acad Sci* 2021;**118**:e2002545117. <https://doi.org/10.1073/pnas.2002545117>
- Joimel S, Schwartz C, Bonfanti J. *et al.* Functional and taxonomic diversity of Collembola as complementary tools to assess land use effects on soils biodiversity. *Front Ecol Evol* 2021;**9**:630919. <https://doi.org/10.3389/fevo.2021.630919>
- Manu M, Băncilă RI, Birsan CC. *et al.* Soil mite communities (Acari: Mesostigmata) as indicators of urban ecosystems in Bucharest, Romania. *Sci Rep* 2021;**11**:3794. <https://doi.org/10.1038/s41598-021-83417-4>
- Meehan ML, Song Z, Lumley LM. *et al.* Soil mites as bioindicators of disturbance in the boreal forest in northern Alberta, Canada: Testing taxonomic sufficiency at multiple taxonomic levels. *Ecol Indic* 2019;**102**:349–65. <https://doi.org/10.1016/j.ecolind.2019.02.043>
- Nielsen UN. *Soil Fauna Assemblages: Global to Local Scales*. 1st edition. Cambridge: Cambridge University Press, 2019. <https://doi.org/10.1017/9781108123518>
- Nielsen UN, Ayres E, Wall DH. *et al.* Soil biodiversity and carbon cycling: a review and synthesis of studies examining diversity–function relationships. *Eur J Soil Sci* 2011;**62**:105–16. <https://doi.org/10.1111/j.1365-2389.2010.01314.x>
- Oriol T, Pasquet J, Cortet J. Automatic identification of Collembola with deep learning techniques. *Eco Inform* 2024;**81**:102606. <https://doi.org/10.1016/j.ecoinf.2024.102606>
- Potapov A, Bellini B, Chown S. *et al.* Towards a global synthesis of Collembola knowledge – challenges and potential solutions. *Soil Organisms* 2020;**92**:3. <https://doi.org/10.25674/SO92ISS3PP161>
- Pravia A, Barreto C, Ashwood F. *et al.* Oribatid mite taxa and composition associated with temperate habitats in Great Britain. *Appl Soil Ecol* 2025;**215**:106471. <https://doi.org/10.1016/j.apsoil.2025.106471>
- Pyatt DG. Soil classification. In: *Number 68/82/SSN in Forestry Commission Research Information Note*. Edinburgh: Forestry Commission, 1982.
- Querner P, Bruckner A. Combining pitfall traps and soil samples to collect Collembola for site scale biodiversity assessments. *Appl Soil Ecol* 2010;**45**:293–7. <https://doi.org/10.1016/j.apsoil.2010.05.005>
- Radford A, Kim JW, Hallacy C. *et al.* Learning transferable visual models from natural language supervision. *arXiv* 2021. <https://doi.org/10.48550/ARXIV.2103.00020>

- Remelli S, Gruss I, Menta C. Advances in soil arthropod identification: integrating morphological, molecular, and AI-based approaches. *Eur J Soil Biol* 2026;**128**:103812. <https://doi.org/10.1016/j.ejsobi.2026.103812>
- Sandler M, Howard A, Zhu M. *et al.* MobileNetV2: inverted residuals and linear bottlenecks. *arXiv* 2019. <https://doi.org/10.48550/ARXIV.1801.04381>
- Schneider S, Taylor GW, Kremer SC. *et al.* Bulk arthropod abundance, biomass and diversity estimation using deep learning for computer vision. *Methods Ecol Evol* 2022;**13**:346–57. <https://doi.org/10.1111/2041-210X.13769>
- Seniczak A, Seniczak S, Iturrondobeitia JC. *et al.* Mites (Oribatida and Mesostigmata) and vegetation as complementary bioindicators in peatlands. *Sci Total Environ* 2022;**851**:158335. <https://doi.org/10.1016/j.scitotenv.2022.158335>
- Shepherd M, Crotty F. *A Key to the Soil Mites of Britain and Ireland*. Number Test version 11. (Unpublished), 2024.
- Silvan N, Laiho R, Vasander H. Changes in mesofauna abundance in peat soils drained for forestry. *For Ecol Manage* 2000;**133**:127–33. [https://doi.org/10.1016/S0378-1127\(99\)00303-5](https://doi.org/10.1016/S0378-1127(99)00303-5)
- Soil Survey of Scotland Staff. *National Soil Map of Scotland*, 1981. <https://doi.org/10.5281/zenodo.4646890>
- Southwood TRE. *Ecological Methods: With Particular Reference to the Study of Insect Populations*. Netherlands, Dordrecht: Springer, 1978. <https://doi.org/10.1007/978-94-009-5809-8>
- Stevens S, Wu J, Thompson MJ. *et al.* BioCLIP: A vision foundation model for the tree of life. *arXiv* 2023. <https://doi.org/10.48550/ARXIV.2311.18803>
- Sys S, Weißbach S, Jakob L. *et al.* CollemboAI, a macrophotography and computer vision workflow to digitize and characterize samples of soil invertebrate communities preserved in fluid. *Methods Ecol Evol* 2022;**13**:2729–42. <https://doi.org/10.1111/2041-210X.14001>
- Thakur MP, Phillips HRP, Brose U. *et al.* Towards an integrative understanding of soil biodiversity. *Biol Rev* 2020;**95**:350–64. <https://doi.org/10.1111/brev.12567>
- Wadoux AMJ. Artificial intelligence in soil science. *Eur J Soil Sci* 2025;**76**:e70080. <https://doi.org/10.1111/ejss.70080>
- Walter DE, Proctor HC. *Mites: Ecology, Evolution & Behaviour: Life at a Microscale*. Netherlands, Dordrecht: Springer, 2013. <https://doi.org/10.1007/978-94-007-7164-2>
- Wehner K, Heethoff M, Brückner A. Seasonal fluctuation of oribatid mite communities in forest microhabitats. *PeerJ* 2018;**6**:e4863. <https://doi.org/10.7717/peerj.4863>
- Wehrli M, Meyer A, Da Silva v S. *et al.* COLLEMBOT: AI-based counting of Collembola for OECD 232 tests. *Environ Toxicol Chem* 2026;vgag068. <https://doi.org/10.1093/etojnl/vgag068>
- Wenglein R, Lanzrein I, Graf M. *et al.* Changes in soil mesofauna communities with increasing proportions of Douglas-fir and silver fir in European beech forests. *Appl Soil Ecol* 2025;**213**:106328. <https://doi.org/10.1016/j.apsoil.2025.106328>
- Wu T, Ayres E, Bardgett RD. *et al.* Molecular study of worldwide distribution and diversity of soil animals. *Proc Natl Acad Sci* 2011;**108**:17720–5. <https://doi.org/10.1073/pnas.1103824108>
- Yang C-H, Feuer B, Jubery Z. *et al.* BioTrove: a large curated image dataset enabling AI for biodiversity. *arXiv* 2024. <https://doi.org/10.48550/ARXIV.2406.17720>
- Zhuang X, Wang B, Xie Z. *et al.* Deep learning method for Collembola identification using single species and community combinations images. *Soil Ecol Lett* 2025;**7**:250352. <https://doi.org/10.1007/s42832-025-0352-9>