Biology Methods and Protocols, 2024, bpae063
https://doi.org/10.1093/biomethods/bpae063
Advance Access Publication Date: 27 August 2024
Methods Article

Deep learning image analysis for filamentous fungi taxonomic classification: Dealing with small datasets with class imbalance and hierarchical grouping

Stefan Stiller [1,2,3], Juan F. Dueñas [3,4], Stefan Hempel [3,4], Matthias C. Rillig [3,4], Masahiro Ryo [1,2,*]

[1] Leibniz Centre for Agricultural Landscape Research (ZALF), Müncheberg 15374, Germany
[2] Institute of Environmental Sciences, Brandenburg University of Technology Cottbus–Senftenberg (BTU), Cottbus 03046, Germany
[3] Institute of Biology, Freie Universität Berlin, Berlin 14195, Germany
[4] Berlin-Brandenburg Institute of Advanced Biodiversity Research (BBIB), Berlin 14195, Germany

*Correspondence address. Research Platform Data Analysis and Simulation, Leibniz Centre for Agricultural Landscape Research, Müncheberg, Brandenburg, Germany. Tel: +49 (0)33432 82-206; Fax: +49 (0)33432 82-334; E-mail: masahiro.ryo@zalf.de

Abstract

Deep learning applications in taxonomic classification for animals and plants from images have become popular, while those for microorganisms are still lagging behind. Our study investigated the potential of deep learning for the taxonomic classification of hundreds of filamentous fungi from colony images, which is typically a task that requires specialized knowledge. We isolated soil fungi, annotated their taxonomy using standard molecular barcode techniques, and took images of the fungal colonies grown in petri dishes ($n = 606$). We applied a convolutional neural network with multiple training approaches and model architectures to deal with some common issues in ecological datasets: small amounts of data, class imbalance, and hierarchically structured grouping. Model performance was overall low, mainly due to the relatively small dataset, class imbalance, and the high morphological plasticity exhibited by fungal colonies. However, our approach indicates that morphological features like color, patchiness, and colony extension rate could be used for the recognition of fungal colonies at higher taxonomic ranks (i.e. phylum, class, and order). Model explanation implies that image recognition characters appear at different positions within the colony (e.g. outer or inner hyphae) depending on the taxonomic resolution. Our study suggests the potential of deep learning applications for a better understanding of the taxonomy and ecology of filamentous fungi amenable to axenic culturing. Meanwhile, our study also highlights some technical challenges in deep learning image analysis in ecology, highlighting that the domain of applicability of these methods needs to be carefully considered.

Keywords: Convolutional Neural Network (CNN); transfer learning; local interpretable model-agnostic explanations (LIME); microbiology; mycorrhizal fungi

Introduction

Deep learning is an artificial intelligence (AI) toolbox inspired by the human brain to learn patterns from complex data. It is one of today's most rapidly growing technical fields [1]. Deep learning has been used successfully in various scientific fields, including biology and medicine [2]. In ecology, deep learning is used, for instance, for identifying species and classifying animal behavior from camera images, audio recordings, and videos [3].

Taxonomic classification with deep learning image analysis has been demonstrated successfully for many taxa, but applications have been limited to individuals with rigid morphological structures, such as plants, animals, and also fruiting bodies of fungi [4–7]. No study has investigated the potential of the technique to study the taxonomy of filamentous fungi that do not produce macroscopically visible reproductive structures, which constitute the large majority within the kingdom. Indeed, it is largely unknown whether filamentous fungi can be taxonomically classified based solely on the macroscopic morphological features when grown in a pure culture. While it is clear that taxonomic identification of an isolate to the level of species is impossible without comparing characters microscopically (as keys mostly focus on fungal spores and spore-producing structures, e.g. [8]), or even at the molecular level, identification at a coarser taxonomic resolution might be feasible. Indeed, some early diverging lineages of filamentous fungi produce consistent colony morphologies in culture (e.g. Mortierellomycota, [9]). Recent work suggests that hyphal growth speed (which has an impact on colony size in the Petri dish) and the complexity of the hyphal architecture are phylogenetically conserved, at least at the phylum level [10]. Because colony characters such as color, shape, exudation, and sporulation are easily quantifiable morphological traits, we wondered whether annotating the coarser levels of the taxonomic hierarchy to a fungal colony is possible based on images using these attributes.

We used the appearance of "colonies" of fungal individuals grown on Petri dishes, representing a higher level of morphological organization than hyphal traits, to test this idea. Because it is known that pure cultures of the same species can exhibit a staggering variation of morphology depending on the growing

Received: 15 May 2024. **Revised:** 29 July 2024. **Editorial decision:** 07 August 2024. **Accepted:** 26 August 2024

conditions, we employed a standardized culturing protocol to produce a large number of colony images. Our study first examined the application of deep learning models to achieve the taxonomic classification of filamentous fungi from the images of fungal colonies grown on Petri dishes. In particular, we asked (i) up to which taxonomic resolution (from phylum, class, order, family, genus, to species level) a deep learning model can keep accuracy and (ii) from which region of the individual fungal colony the model learns the morphological characteristics for taxonomic classification.

In addition to these questions, we attempted to tackle some technical challenges common in ecological datasets. The size of a dataset is usually small (10^2–10^3) since experimental tasks are labor intensive, generally resulting in low model performance. The data containing missing values are not easy to impute in a meaningful way (e.g., many fungi cannot be identified at species level). The probability distribution of categories is highly imbalanced (e.g. many photo images belong to a handful of taxonomic groups). The classification property is hierarchically nested (e.g. once the species name is known, its higher taxonomic names are determined because of phylogeny). Since these data properties are pretty standard in ecological datasets, we think that overcoming some of the technical challenges is valuable for exploring the potential application of deep learning image classification in ecology in a broader context.

Methods and materials

Dataset: FunTrait image collection

We obtained a subset of fungal culture images from a culture collection isolated from the German Biodiversity Exploratories' very intensive research plots (VIPs, [11]). These are located on grasslands in three different areas across Germany (nine per site, $n = 27$). VIPs represent a replicated gradient of grassland management intensity that ranges from near-pristine to heavily managed for agricultural purposes. A total of 500 g of soil from the upper 10 cm beneath the surface (e.g. [12]) were collected per plot. Each sample represents a composite of five random soil cores across each plot. Upon collection, soil samples were transported to the Institute of Biology at Freie Universität Berlin, where they were stored in the dark at 4°C.

Using an in-house high-throughput culturing approach, we obtained fungal colonies from soil samples. All work, from dilution to colony isolation, was done under sterile conditions. Briefly, the soil was subjected to serial dilutions to hinder the recovery of highly sporulating fungi. Several morphologically contrasting fungal isolates were recovered from each diluted soil solution using potato dextrose agar (PDA, X931.2. Roth) in full and 1/10 strength, incubation at 12°C and subsequent transfer on new agar plates.

The cultures used to produce images grew for approximately 10–20 days at ambient temperature (∼15–20°C) in 6-cm diameter Petri dishes plated with 100% (w/v) sterilized malt extract agar (MEA, X923.1 Roth). To hinder bacterial growth, we mixed the sterilized MEA solution with a range of antibiotics (Penicillin G 24 μg/l, Chlortetracycline 48 μg/l, and Streptomycin 26 μg/l) before casting.

Each isolate was then taxonomically annotated via DNA extraction and Sanger sequencing. Before sequencing, the entire extent of the internal transcribed spacer (ITS) and large subunit (LSU) regions within the recombinant DNA (rDNA) operon was amplified by employing the ITS1 [13] and NL4 [14] primers. The ITS variable regions were then extracted *in silico* with ITSx [15]. Finally, the blast+ classification algorithm [16] was employed to query the ITS sequences of each isolate to UNITE's reference database from February 2020 [17]. Because UNITE databases employ the taxonomic framework proposed by [18], taxonomic identities inherited by the query sequences follow this framework. Each isolate was identified to the finest taxonomic level possible.

Because the goal of image identification was ancillary to the main objective of the fungal collection, Petri dishes were arranged in a semi-standardized fashion to obtain a group image. Images were group scans of up to 12 Petri dishes over a blue background screen (Fig. 1). Plates were distributed in a 3 × 4 matrix such that their position on the image corresponded to the position of their DNA extract on a 96-well DNA extraction plate. It allows the correspondence of an ITS sequence with an individual image of an isolate. Petri dishes were scanned with a Perfection V800 scanner (Seiko Epson Corporation, Japan) from the bottom and top. At the bottom of each plate, a handwritten mark with the isolate code was present. The images were formatted as file types .tif, .jpg, or .png and in resolutions 2550 × 3509, 5100 × 7019, or 6800 × 9359, respectively. Only top-view images were included. The dataset was composed of 606 fungal isolates as of 18 March 2020, when we established our initial study (Fig. 2).

Biological surveys inherently suffer from class imbalance, meaning that only a few taxonomic groups are highly abundant within the dataset. In the present case, molecular methods classified isolates into 5 phyla (Ascomycota, Basidiomycota, Mucoromycota, Mortierellomycota, and Zoopagomycota), 11 classes, and 190 species (Fig. 3a). Because of the hierarchical organization of taxonomy, class imbalance propagates throughout the ranks, such that an imbalance occurring at a given taxonomic rank is affected by the imbalance of the preceding rank (Fig. 3b–g). Another problem of taxonomic classification by molecular methods is missing values. Missing values can occur when an isolate can be assigned to a higher taxonomic

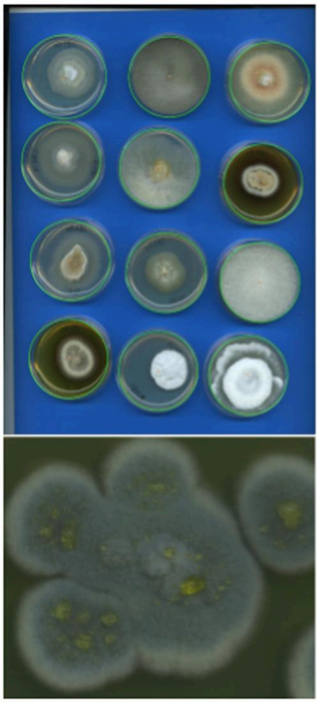

Figure 1. Example of a scanned image of cultures in the dataset (upper). Circles denote Petri dish object detection using the Circle Hough Transform. A zoom-in on the bottom-left sample as an example of a filamentous fungus (lower).

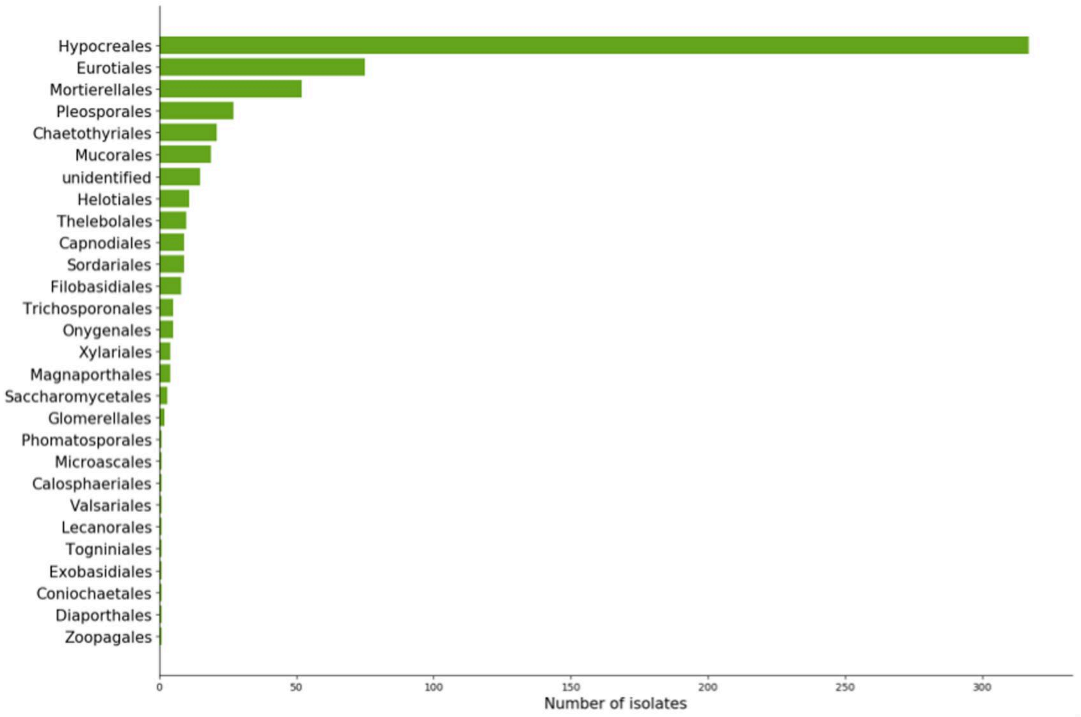

Figure 2. Absolute number of isolates assigned to a known order by DNA extraction and Sanger sequencing method. Fungal isolates were extracted from soil samples, and the cultures were bred in Petri dishes. Taxonomic identification was annotated using blast+ classification algorithm, querying isolate ITS region to the UNITE database.

level (e.g. class) but not reliably to any of the lower taxonomic levels (order, family, etc.). In this dataset, we had missing values for 11 instances for phylum, 39 for class, 40 for order, 67 for family, 119 for genus, and 319 for species. We replaced the missing values by taking the taxonomic name of the higher rank with a suffix indicating the current rank (e.g. for Ascomycota, Ascomycota_class, …, Ascomycota_species), as we were concerned that the missing values might introduce a significant bias.

Image preprocessing

At first, the group image was cut, such that each sub-image contained one Petri dish scan. The images were resized to 224 × 224 pixels using pixel area relation to reduce the computational burden (except for augmented random oversampling to 356 × 356 pixels; see below). We used the Circle Hough Transform to automatically detect all Petri dishes' positions in a single image. We used the implementation by the Open CV Project (*OpenCV* 4.4.0), which uses Hough Gradient [19] (Fig. 1). Then, each identified sample was annotated by a set of six hierarchically nested labels, each representing a taxon given a taxonomic rank, which is phylum > class > order > family > genus > species.

Three datasets for model training

We prepared three datasets for model training to deal with class imbalance: (i) the original data as preprocessed above; (ii) naive random oversampling; and (iii) augmented random oversampling. These methods aimed to ensure that the neural network does not become biased toward the more common classes by providing a more balanced representation of all classes during training.

Original data (without resampling)

The dataset contained images of fungal colonies without any modifications. The images were composed of 224 × 224 pixels. Each image contains a single individual that is taxonomically annotated.

Naive random oversampling

This method artificially increased the representation of underrepresented classes by duplicating and slightly altering their images, such as flipping or adjusting lighting, to balance the dataset. One approach to class imbalance is to use under- and oversampling to equalize the class distributions [20]. In undersampling, samples are randomly drawn from the original set down to a lower limit. Oversampling is its counterpart, where samples are randomly drawn and re-added to the setup up to a higher limit. Undersampling is not the approach of choice due to the already small size of the dataset. The problem would almost become a one-shot learning problem, which is increasingly difficult to solve. In naive random oversampling, the distribution is balanced out by increasing the sample size in the minority classes by adding slightly altered samples drawn from the minority classes while retaining the taxonomic annotation of the donor. For image alteration, the samples have their axes randomly flipped and their lighting adjusted.

Augmented random oversampling

This technique not only duplicated and altered images like naive oversampling but also included more complex transformations like resizing and cropping to further enhance the diversity of the augmented images. Augmented random oversampling applied naive random oversampling and additionally two image augmentation techniques. The first augmentation was done at the image preprocessing stage. The high-resolution images were not resized to 224 × 224 pixels but 356 × 356 pixels. Then, from a 356 × 356 pixel image, multiple 224 × 224 pixel images were generated by cropping the image at random positions. It can be interpreted as

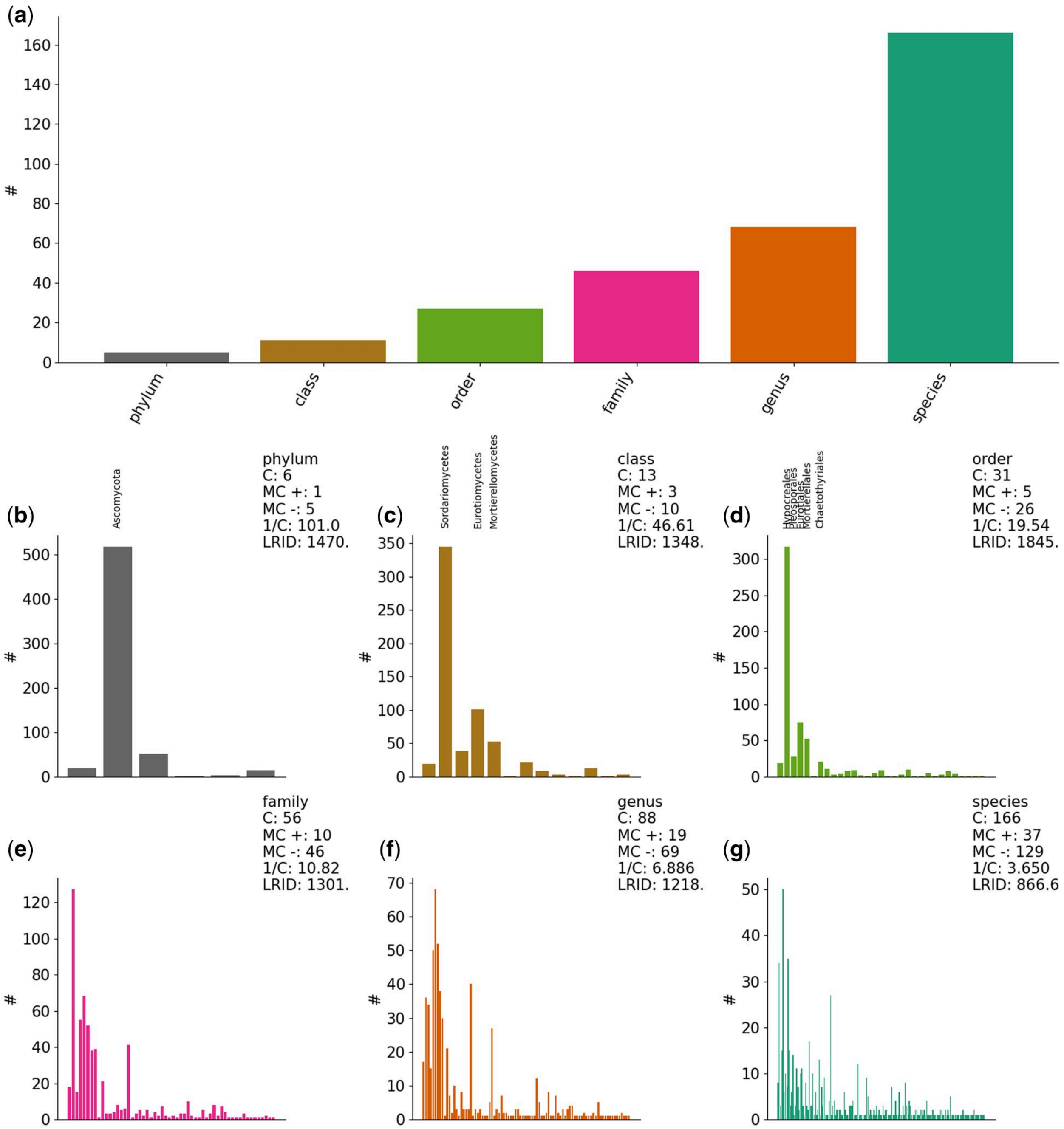

Figure 3. (**a**) Number of unique taxa/classes according to taxonomic rank for 606 samples. (**b–g**) Frequency distribution within each taxonomic rank. C is the number of categories, MC+ is the majority class count, MC− minority class count. A category is regarded as majority if the number of samples 'hash' is higher than the average number of samples per category. LriD is the likelihood ratio imbalance degree. For phylum, class, and order levels, the majority of classes are labeled for visualization.

zooming into a part of the image. The second augmentation was conducted on the images by applying a color jitter concerning brightness, contrast, and saturation.

Modeling with deep learning

Algorithm

We applied a convolutional neural network algorithm, DenseNet-169 [21], with transfer learning to deal with the small sample size. DenseNet has some compelling technical advantages: the method is robust to the vanishing-gradient problem, strengthens feature propagation, encourages feature reuse, and substantially reduces the number of parameters. We applied transfer learning [22]. Transfer learning is a technique that uses a model that is already pre-trained with a vast dataset to make a model efficient. Pre-trained models were taken from MXNET Gluon Model Zoo (MXNET Gluon Model Zoo. Classification–gluoncv 0.11.0 documentation).

Model architecture

We tried three different model (classifier) architectures run with the prepared three datasets: Separate Local (SL) per-level classifiers, Multi-Label (ML) classifiers, and Hierarchically Chained (HC) Local per-level classifiers. These approaches explored different ways to incorporate or bypass the hierarchical structure inherent in taxonomic classification, allowing us to assess their effectiveness in a deep learning context. Silla and Freitas [23] suggest several concepts to engage hierarchical classification, and we applied three of them in this study.

SL was the simplest approach where each taxonomic rank has its own independently trained classifier. This method treats each taxonomic level separately, without sharing information across levels, which means that each classifier operates independently of the others, potentially missing hierarchical patterns.

ML was a single, relatively complex model that considered the entire taxonomic hierarchy in one go [23]. This global classifier can potentially learn and leverage the hierarchical relationships within the data, but because the output units for each rank are independent, it might produce biologically inconsistent classifications.

HC was the most complex model in this study, where each classifier for a taxonomic rank passes its learned features to the next level in the hierarchy. This approach directly utilized the hierarchical structure, with each level's classifier influenced by the outputs from higher levels, theoretically improving consistency and accuracy in classification. Based on taxonomic rank, a deep neural network classifier learns to discriminate taxons. Then, the learned parameters are passed on to the hierarchically nested successive rank classifier, making use of transfer learning. HC makes direct use of the taxonomic hierarchy information with six hierarchically stacked classifiers $C_{\mathrm{Phylum}} < C_{\mathrm{Class}} < C_{\mathrm{Order}} < C_{\mathrm{Family}} < C_{\mathrm{Genus}} < C_{\mathrm{Species}}$. Each classifier C has its nested loss function J_{Rank} with respect to hierarchy, s.t. $J_{\mathrm{Species}} = J_{\mathrm{Species}}(J_{\mathrm{Genus}}(J_{\mathrm{Family}}(J_{\mathrm{Order}}(J_{\mathrm{Class}}(J_{\mathrm{Phylum}})))))$. Six HC models apply one output unit each. Those units are in the sense that the preceding level's output directly influences this level's units.

The model architectures with hyperparameter settings were as follows. As specified by Huang *et al.* [21], DenseNet-169 utilizes rectified linear units [24], Batch Normalization [25], and pooling [26]. We applied L2 regularization [27] with weight decay $wd = 0.01$, learning rate $lr = 0.001$, and momentum $m = 0.8$. Parameter initialization follows the approach [28], Xavier initialization. We trained the models using a softmax cross-entropy loss function [29].

Model training and performance evaluation

Models are trained to maximize their performances, making the selection of appropriate performance indicators crucial for assessing their effectiveness. We used the Accuracy and Matthews correlation coefficient (MCC). These metrics help in evaluating how well the model predicts the correct taxonomic classifications, providing a quantitative measure of the model's accuracy and reliability.

Accuracy is the most widely used indicator for classification tasks, measuring the proportion of correct predictions among the total number of cases. However, it can be misleading in datasets with class imbalance, as it may still yield high scores by predominantly predicting the majority classes correctly while ignoring the minority classes.

MCC addresses the limitations of Accuracy, particularly in imbalanced datasets [30]. MCC provides a comprehensive measure of classification quality, accounting for true positives, true negatives, false positives, and false negatives, making it a more reliable indicator of model performance when classes are not equally represented. The MCC indicator was initially proposed by Matthews [31] and recently regained the attention of the deep learning community. An MCC score of 1 indicates perfect prediction, 0 indicates no better than random prediction, and negative values indicate worse-than-random performance. Therefore, we think MCC is a more honest indicator for evaluating model performance under class imbalance than Accuracy.

Model performance was tested on data that were simply split into a train and test part according to a ratio of 70% for training data and 30% for test data. In addition to the model performance assessment, we further tested the robustness of the model performance using 5-fold cross-validation. The data were split into five parts of equal size. In five iterations, in alternating order, one such part was denoted as test data, while the others accumulated to the training data. Although typical image classification studies do not employ cross-validation, we considered this technique important for small datasets to quantitatively assess the stability of model estimates influenced by the number of samples.

Opening the black-box model: explanation

Explaining why a black-box model made a prediction is important to understanding how a prediction is made [32, 33]. An untrustworthy model could make a correct prediction based on an inappropriate reason: For instance, in this study, it is possible that the model performs well, but what it learned is not a fungal colony-level trait but the handwritten time stamp on the Petri dish. One way to explain deep learning models is to use model-agnostic explanations such as local interpretable model-agnostic explanations (LIME) [34]. LIME is a post-hoc, local surrogate model that is interpretable and can explain individual predictions.

We applied LIME for some samples that revealed a high score to inspect what the models learned visually. We set the following hyperparameters: kernel size = 6, max distance = 50, ratio = 0.5, neighborhood size = 1000, and selected features = 100. We applied a standard technique for LIME, namely, superpixel [35], with the quick-shift algorithm [36].

Settings

All models were built in Python 3.6.9 using Apache's MXNET gluon framework with GPU support CUDA-10.1. All deep learning computations were run on Google Colab. LIME 0.2.0.1 was run locally on a workstation in an Ubuntu Bionic Beaver environment. Each model was trained in 20 epochs of fine-tuning on the three datasets.

Results

At the training phase, all model and dataset combinations reached a stabilized performance score (both Accuracy and MCC) after at most 20 epochs of parameter-tuning, that is successive model training over 20 times using the whole dataset, indicating that the duration of the training period was satisfactory (Fig. 4 as a representative case; see Supplementary Fig. S1, S2, S3 for the others). At the testing phase, all model and dataset combinations showed a substantial drop in performance (both Accuracy and MCC), meaning that the models overfitted and learned false data characteristics, which are irrelevant for taxonomic group prediction. For example, in Fig. 4, we can observe that the model successfully learned from the training data as depicted as increasing performance in the upper curves. However, what it learned was not general enough or not fitting to the task to identify unseen samples with high accuracy (cf decreased test performance shown in the lower curves).

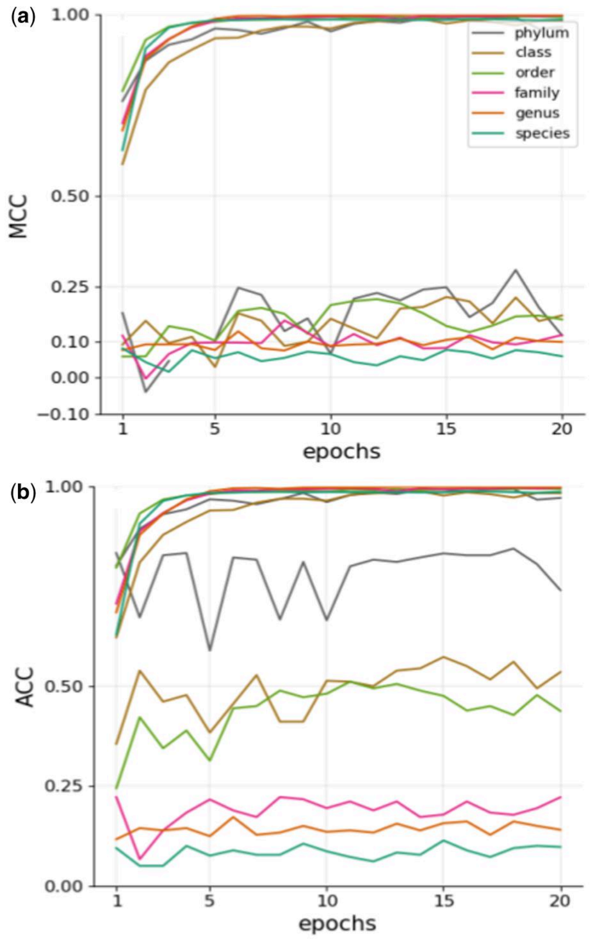

Figure 4. Performance for SL per-level classifiers finetuned in 20 epochs according to MCC on (**a**) original, (**b**) naive oversampled, (**c**) transform oversampled data and according to Accuracy on (**d**) original, (**e**) naive oversampled, (**f**) transform oversampled datasets.

While reducing performance in the test phase, all model–dataset combinations at the best epoch kept non-zero MCC scores, and most of them fell into the range of 0.1–0.2, meaning that the models are better than random (Fig. 5a and b). Performance scores changed along with taxonomic rank. Accuracy, as well as MCC, showed a steady decline trend from the phylum to species level. As a general trend, the SL model architecture outperformed ML and HC architectures, and HC was no better than ML. With cross-validation, all models reduced MCC scores by about 0.1 (Supplementary Figs S4 and S5), but they still showed non-zero scores.

Focusing on specific model–dataset combinations based on MCC, we observed the following. The SL with naive random oversampling performed best in three out of the six taxonomic ranks. Following the three SL models, the HC model with the original dataset was placed at the fourth performance. However, the other two HC models were no better than the ML models.

Accuracy scores reached >0.75 in many model trials at the phylum level but then linearly declined along with taxonomic resolution, reaching about 0.1. However, these high Accuracy scores, particularly at the phylum level, are misleading due to the strong class imbalance, as they do not account for the low representation of minority classes (Fig. 3). This issue is further evidenced by the modest MCC scores, which provide a more balanced view of model performance by considering true positives, false positives, true negatives, and false negatives. The MCC scores remained low, indicating that while the model performed well on the majority classes, it struggled with minority classes. The two approaches to alleviate class imbalance, naive random oversampling and augmented random oversampling, did not show substantial improvement over the original dataset, with differences in MCC scores being minor (less than 0.05).

We found that the models learned a few taxonomic groups better than others, as shown for the SL classifier in Fig. 6 (for other classifiers, see Supplementary Figs S6 and S7). At the class level, Sordariomycetes and Eurotiomycetes were well predicted, and at the order level, Hypocreales and Eurotiales were relatively well predicted. The classifier predicted with an Accuracy of 0.83 for phylum, 0.56 for class, and 0.49 for order, and taking class imbalance into account, 0.30, 0.22, and 0.21 MCC, respectively. This result indicates that not all but some fungal taxa display unique morphological characteristics at the colony level. On the other hand, this result might reflect the fact that a large majority of the isolates annotated as Hypocreales corresponded to the species *Clonostachys intermedia*, which probably exhibits a fairly consistent colony morphology.

To roughly grasp the explanations that led to the model's prediction, we took some specific examples and applied LIME for images. We selected the ML model with naive random oversampling at the best-performing epoch because of the intrinsic model architecture, which can learn colony traits important for classification that translate throughout taxonomic ranks. We looked at explanations for correctly classified individuals with at least five out of six taxonomic ranks. We observed that the explanation highlights a complete outline for phylum and class. Given a Petri dish surface cover after a specific time of fungal growth, a full outline could be interpreted as hyphal growth speed (see model explanations Fig. 7 for *Apiotrichum dulcitum*, and Supplementary Fig. S8 for *Penicillium araracuarense*). In addition, some model explanations shown in Figure 7 point to the agar medium, as well as to the rim of the Petri dish. These could indicate that the model looked for handwriting at the position in the agar medium, since few dishes contained writing, or used the position of the dish at the time of photo taking (i.e. size and degree of the visible edge). This demonstrates the models' sensitivity to the slightest differences in photo taking. In lower taxonomic ranks, the outline thinned. Moreover, specific structures in the fungal colony surface were highlighted. We observed that the explanation of fungi with more complex surfaces is more challenging to interpret (Supplementary Fig. S8). However, explanations seem to cluster around colony structures, too.

Discussion

In this work, we explored the application of deep learning on taxonomically labeled image data of colonies of filamentous fungi and explanations behind the classification. We demonstrated that deep learning could classify some isolates at least at the phylum, class, and order levels, regardless of the diverse nature of their appearance and the difficulties inherent to the dataset. Evidently, the outlines and inner regions of fungal colonies contributed to high prediction scores, which might represent the macroscopic expression of hyphal growth and structuring. The higher prediction scores for phylum, class, and order achieved, and the visual explanations of the predictions at the outlines and inner regions of the fungal colonies align with phylogenetic conservation of morphological traits observed previously [10]. However, additional model explanations in the agar medium and

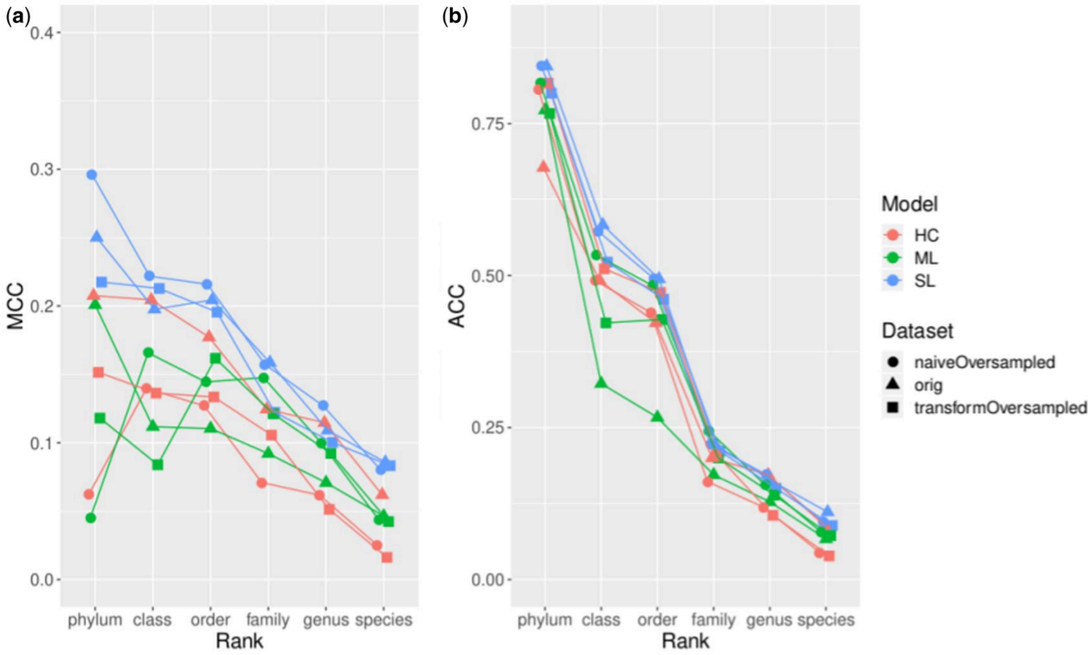

Figure 5. Comparison of best test performances each model achieved according to (**a**) MCC and (**b**) Accuracy on 606 samples.

at the rim of the Petri dish indicated that the absence or presence of handwriting and the photo taking angle may have contributed to the classification task. This raises the question of which importance these three explanation regions contributed to classification.

Highly skewed data

The sample distribution was highly skewed, and the models were able to predict samples that made up the majority of the dataset but failed to predict most minority groups despite class imbalance mitigation approaches. Nonetheless, we found that all models performed best on the naive oversampled dataset regarding performance and stability. It accelerated learning and seemed to boost test scores for phylum, class, and order while also reducing the noise's influence. Augmented oversampled data also accelerated learning, yet at a slower pace. However, the general classification score for this set on the test data dropped dramatically, which might indicate that color is an essential feature for taxonomic identification.

We also showed that MCC is more robust than Accuracy for evaluating the performance of deep learning models (or, more generally, any multi-classification problems in ecology). As we demonstrated, Accuracy is not a plausible indicator when the data are highly imbalanced, which is a common issue in ecology (e.g. a few species dominate the majority of relative abundance in a community). Instead, MCC is a fairer, more honest indicator for assessing model performance. An alternative is the Synthetic Minority Over-sampling Technique [37]. It offers a logical way of the synthetic creation of new samples in similarity to the existing ones.

At the class level, Sordariomycetes were well predicted. This can be attributed to the large frequency at which we recovered Clonostachys intermedia ($n > 150$) in our culture set. This means that from a total of approximately 300 isolates that were classed as Hypocreales, half belong to this species only. This unique morphological feature was probably captured during model training.

Hierarchical modeling

Tackling the taxonomy's hierarchical nature, we found that SL per-level classifiers reached a slightly higher performance score than the other models. Unexpectedly, HC classifiers achieved the worst performance. A possible reason is that the hierarchical nature of taxonomic data was not fully leveraged by the current model design or that employing overly complex classifiers increases the chance of error propagation, rather than increasing general performance. This performance can be improved by increasing the sample size or adjusting the model design. Another possible and more ecologically exciting reason is that the inclusion of the hierarchical nature does not improve because the macroscopic morphological characteristics of fungi are phylogenetically not conserved along with taxonomic ranks. The fungal kingdom is full of examples where homologous morphological manifestations have appeared independently within disparate lineages. One example of this is the repeated transition from filamentous to yeast morphologies among distant fungal clades [38]. Another example is the existence of several pleomorphic species within the kingdom [39]. With the wide adoption of molecular phylogenetics, and more recently phylogenomics, fungal taxonomists have confirmed how misleading and inconsistent morphological-based classifications can be, hence the calls for the modernization of fungal systematics [40]. On the other hand, it must be noted that fungal phylogenies are far from being clearly resolved [9]. Hence the use of an imperfect hierarchical classification to train a model has to be viewed with caution. We consider, nevertheless, that ML is the most promising approach in terms of the balance between performance and possible self-learned inclusion of hierarchy.

Small dataset

The small dataset ($n = 606$) significantly limited the deep learning model's performance and generalizability—a common challenge for ecology. Yet, deep learning models show superior performance for many tasks due to their ability for inherent feature extraction. To address the challenge of the small sample size, we

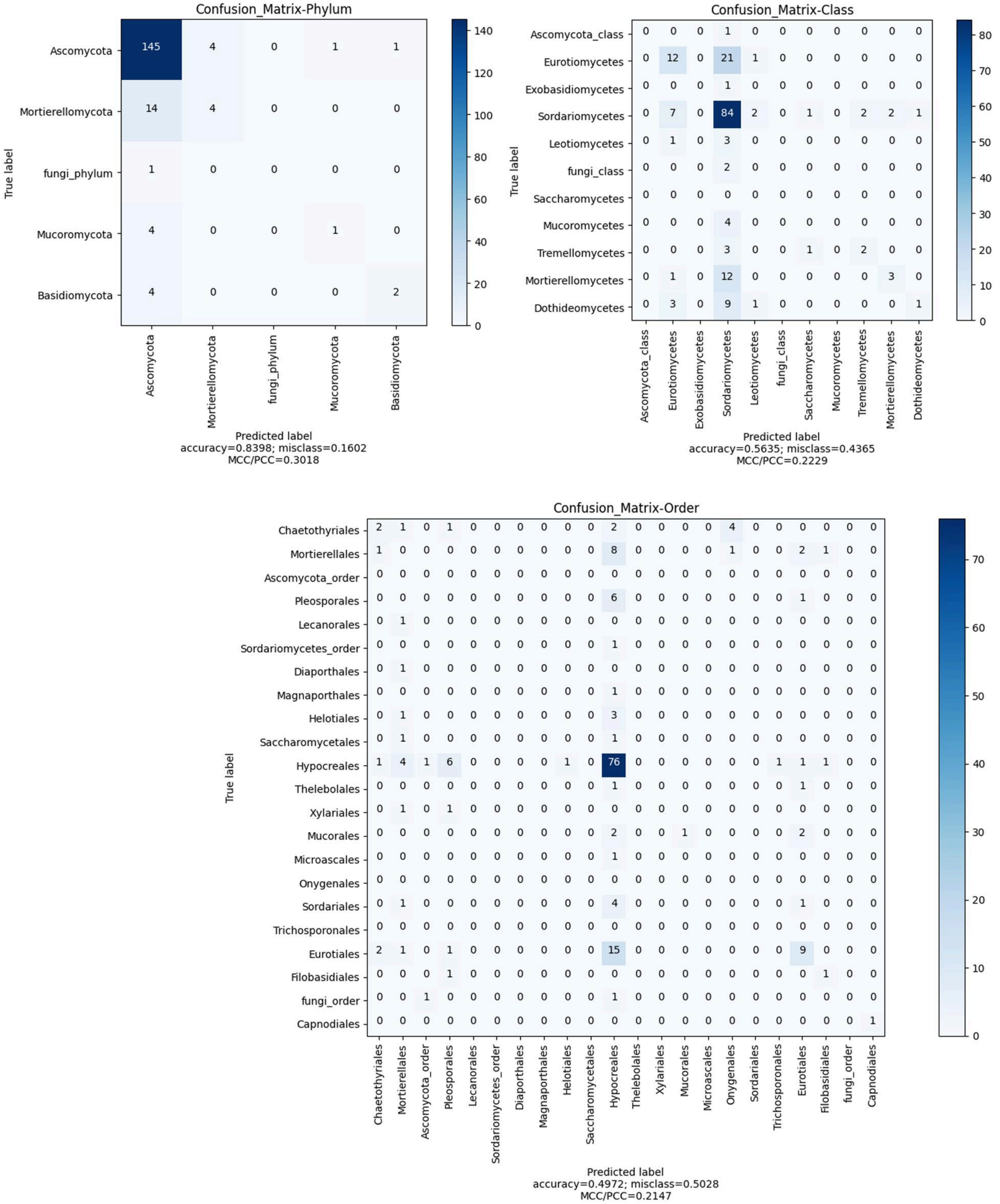

Figure 6. Confusion matrix of SL per-level classifiers trained on 606 samples of naïve oversampled dataset, showing observed versus predicted taxonomic group. Prediction on test data for taxonomic ranks (**a**) phylum at epoch 17, (**b**) class at epoch 14, and (**c**) order at epoch 11.

applied transfer learning and also checked performance stability with cross-validation. Cross-validation is rarely applied for deep learning approaches, but it can be done thanks to the small sample size. Given this, we could carefully check if the performance was just by chance or not. We think this is a more honest approach than showing a point estimate of a performance indicator. However, it has a clear drawback, since the sample size becomes even smaller by keeping a part of the data for validation. We conducted the same analyses after obtaining additional ca. 300 samples, but we observed no qualitative improvement from the 606 samples we investigated (see Supplementary Figs S9–S12). One possible approach to compensate for the lack of data is to add predictors that can represent some critical ecological information. In our case, for instance, technically it is possible

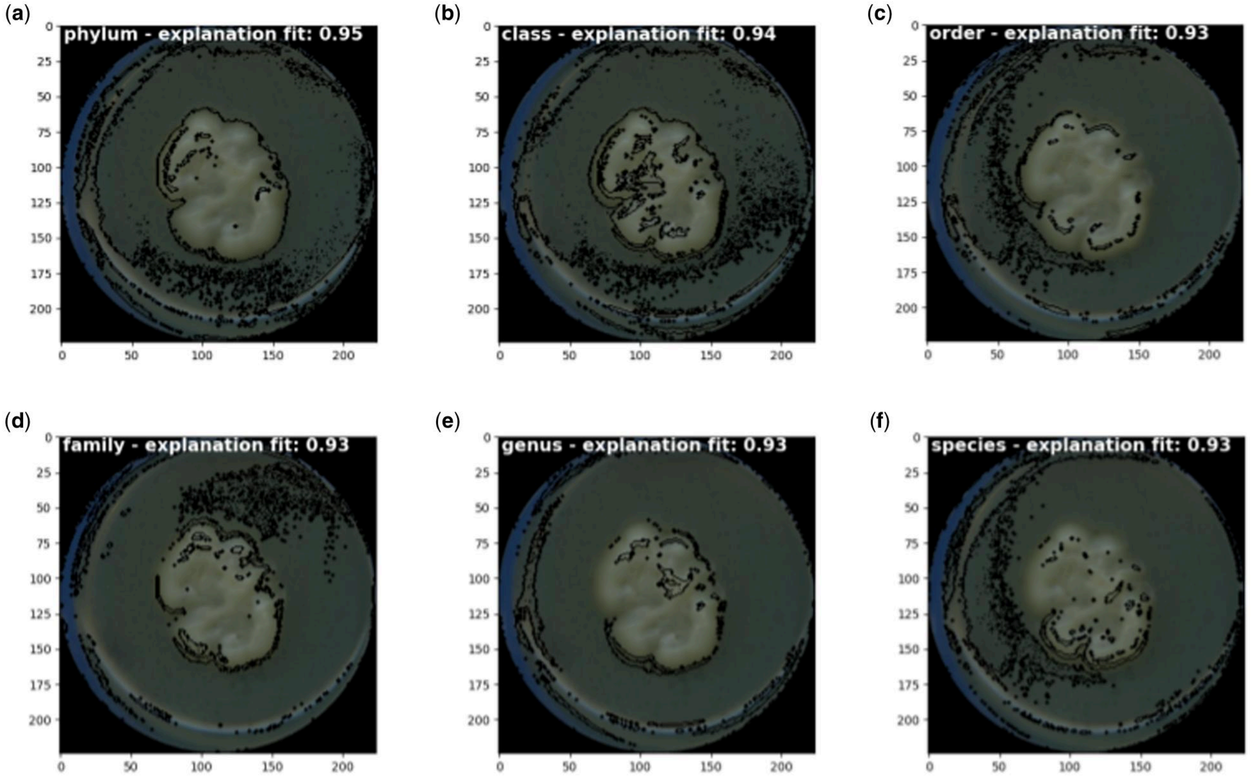

Figure 7. The LIME explanation for *Apiotrichum dulcitum* predicted with ML at epoch 4, with neighborhood size 1000 and 100 superpixels. Segmentation is performed by quickshift algorithm with kernel size 6, max distance 50, and ratio 0.5. Black highlighted areas are LIME explanations at rank (**a**) phylum, (**b**) class, (**c**) order, (**d**) family, (**e**) genus, and (**f**) species.

to add the information about the period of growth time as a predictor in addition to the images, which may help the deep learning to reflect hyphal growth speed as key information that potentially improves the model performance.

Limitations

Altogether, in our study, we faced several limitations that are common for the data-driven analysis in ecology, including highly skewed data, hierarchical dependencies, and small data size, which resulted in over-fitting the training data and limited generalization ability. Addressing these challenges by increasing the sample size, reducing class imbalance, and preventing overfitting is sometimes not feasible logistically, thus posing limitations to the analysis. Additionally, the low predictive power of our models, as indicated by modest MCC scores ranging from 0.1 to 0.2, underscores these challenges and the high morphological variability in fungal colonies. Despite these challenges, deep learning modeling offers benefits such as automatic feature extraction, handling of high dimensional data, and capturing of spatial hierarchies that can be important for many ecological studies [3]. Hence, we propose facilitating data analysis by using approaches that mitigate small data limitations [41], such as image augmentations, transfer or self-supervised learning, and cross-validation. Moreover, Tendle and Hasan showed that the use of self-supervision for model training helps to address overfitting [42].

Furthermore, complex deep learning models can be regarded as black-box models and lack explanations. In addition to achieving high prediction performance, adding model explanations that highlight image regions important for prediction is crucial. For example, we found that the model used several regions that were not part of the fungal colony for predicting by using LIME [34]. How much each region contributed to prediction performance remained unanswered. Future studies can make use of methods like Grad-CAM [43] to address these questions.

Outlook

Deep learning has proven its potential for classification analysis of image data and has found its way into ecology for different ecosystem studies and scales [3]. To our knowledge, we were the first to apply an explainable artificial intelligence technique to mycelia of filamentous fungi. Improving predictive quality should be a priority for follow-up studies. This could be achieved by utilizing higher-resolution images and by increasing the amount of data the algorithm can learn from while keeping class imbalance low. For example, this technique can be applied to microscopic images. Furthermore, studies that visually explore phylogenetic conservation of mycelium morphological traits can synthesize the deep learning model's intrinsic hierarchical structure with the hierarchical structure of taxonomic data.

Acknowledgements

We thank the managers of the three Exploratories, Kirsten Reichel-Jung, Iris Steitz, and Sandra Weithmann, as well as Juliane Vogt, Miriam Teuscher, and all former managers, for their work in maintaining the plot and project infrastructure; we are also grateful to Christiane Fischer for giving support through the central office, Andreas Ostrowski for managing the central database, and Markus Fischer, Eduard Linsenmair, Dominik Hessenmöller, Daniel Prati, Ingo Schöning, François Buscot, Ernst-Detlef Schulze, Wolfgang W. Weisser, and the late Elisabeth Kalko for their role in setting up the Biodiversity

Exploratories project. We thank the administration of the Hainich National Park, the UNESCO Biosphere Reserve Swabian Alb, and the UNESCO Biosphere Reserve Schorfheide-Chorin, as well as all landowners for the excellent collaboration. Field work permits were issued by the responsible state environmental offices of Baden-Württemberg, Thüringen, and Brandenburg. S.H. acknowledges support from Moritz Mittelbach and our student helpers Eka Giorgashvilli and Orestis Kanaris.

Author contributions

Stefan Stiller (Conceptualization [supporting], Formal analysis [lead], Methodology [lead], Software [lead], Validation [lead], Visualization [lead], Writing—original draft [lead], Writing—review & editing [supporting]), Juan F. Dueñas (Data curation [lead], Investigation [equal], Resources [lead], Writing—review & editing [equal]), Stefan Hempel (Data curation [supporting], Funding acquisition [lead], Project administration [equal], Resources [lead]), Matthias C. Rillig (Funding acquisition [supporting], Project administration [supporting], Resources [supporting], Supervision [supporting], Writing—review & editing [supporting])

Supplementary data

Supplementary data is available at *Biology Methods and Protocols* online.

Conflict of interest statement. We do not have conflicts of interest.

Funding

The work has been partly funded by the Deutsche Forschungsgemeinschaft (DFG) Priority Program 1374 "Biodiversity-Exploratories" (HE 6183/5-1).

Data availability

Data is available upon request.

References

1. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. *Science* 2015;**349**:255–60.
2. Ching T, Himmelstein DS, Beaulieu-Jones BK *et al.* Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface* 2018;**15**:20170387.
3. Christin S, Hervet É, Lecomte N. Applications for deep learning in ecology. *bioRxiv*. 2018;334854.
4. Miao Z, Gaynor KM, Wang J *et al.* Insights and approaches using deep learning to classify wildlife. *Sci Rep* 2019;**9**:8137.
5. Seeland M, Rzanny M, Boho D *et al.* Image-based classification of plant genus and family for trained and untrained plant species. *BMC Bioinformatics* 2019;**20**:4.
6. Hansen OLP, Svenning J-C, Olsen K *et al.* Species-level image classification with convolutional neural network enables insect identification from habitus images. *Ecol Evol* 2020;**10**:737–47.
7. Dong JE, Zhang J, Zuo ZT *et al.* Deep learning for species identification of bolete mushrooms with two-dimensional correlation spectral (2DCOS) images. *Spectrochim Acta A Mol Biomol Spectrosc* 2021;**249**:119211.
8. Domsch K. *Compendium of Soil Fungi. [Taxonomically revised by Walter Gams]*. Eching, Germany: IHW, 2007, 672.
9. Naranjo-Ortiz MA, Gabaldón T. Fungal evolution: diversity, taxonomy and phylogeny of the Fungi. *Biol Rev Camb Philos Soc* 2019; **94**:2101–37.
10. Lehmann A, Zheng W, Ryo M *et al.* Fungal traits important for soil aggregation. *Front Microbiol* 2019;**10**:2904.
11. Fischer M, Bossdorf O, Gockel S *et al.* Implementing large-scale and long-term functional biodiversity research: the Biodiversity Exploratories. *Basic Appl Ecol* 2010;**11**:473–85.
12. Vályi K, Rillig MC, Hempel S. Land-use intensity and host plant identity interactively shape communities of arbuscular mycorrhizal fungi in roots of grassland plants. *New Phytol* 2015; **205**:1577–86.
13. Gardes M, Bruns TD. ITS primers with enhanced specificity for basidiomycetes—application to the identification of mycorrhizae and rusts. *Mol Ecol* 1993;**2**:113–8.
14. O'Donnell K. Fusarium and its near relatives. In: Reynolds DD (ed.), *The Fungal Holomorph: Mitotic, Meiotic and Pleiomorphic Speciation in Fungal Systematics* CAB, United Kingdom: Wallingford, 1993, 226.
15. Bengtsson-Palme J, Ryberg M, Hartmann M *et al.* Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi and other eukaryotes for analysis of environmental sequencing data. *Methods Ecol Evol* 2013;**4**:914–9.
16. Camacho C, Coulouris G, Avagyan V *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* 2009;**10**:421.
17. Abarenkov K, Zirk A, Piirmann T *et al. UNITE General FASTA Release for Fungi*. UNITE Community. Version 04.02. 2020. https://unite.ut.ee/repository.php (5 August 2021, date last accessed).
18. Tedersoo L, Sánchez-Ramírez S, Kõljalg U *et al.* High-level classification of the Fungi and a tool for evolutionary ecological analyses. *Fungal Divers* 2018;**90**:135–59.
19. Petković T, Lončarić S. An extension to Hough transform based on gradient orientation. *arXiv:151004863* [cs] [Internet]. 16 October 2015 [cited 19 June 2021], http://arxiv.org/abs/1510.04863 (19 June 2021, date last accessed).
20. Batista G, Bazzan A, Monard MC. Balancing *Training Data for Automated Annotation of Keywords: a Case Study* (undefined) [Internet]. 2003 [cited 19 June 2021], https://www.semanticscholar.org/paper/Balancing-Training-Data-for-Automated-Annotation-of-Batista-Bazzan/c1a95197e15fa99f55cd0cb2ee14d2f02699a919 (19 June 2021, date last accessed).
21. Huang G, Liu Z, Van Der Maaten L *et al.* Densely connected convolutional networks. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. New York, NY, USA: Institute of Electrical and Electronics Engineers, 21–26 July 2017, 2261–9.
22. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. Massachusetts, USA: MIT Press, 2016. [Database]
23. Silla CN, Freitas AA. A survey of hierarchical classification across different application domains. *Data Min Knowl Disc* 2011; **22**:31–72.
24. Glorot X, Bordes A, Bengio Y. Deep Sparse Rectifier Neural Networks. In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. Fort Lauderdale, FL, USA, 11–13 April 2011. PMLR 2011;**15**:315–23.
25. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. *arXiv:150203167 [cs] [Internet]*. 2015 March 2 [cited 19 June 2021], http://arxiv.org/abs/1502.03167 (19 June 2021, date last accessed).
26. Lecun Y, Bottou L, Bengio Y *et al.* Gradient-based learning applied to document recognition. *Proc IEEE* 1998;**86**:2278–324.
27. Ng AY. Feature selection, L_1 vs. L_2 regularization, and rotational invariance. In: *Proceedings of the twenty-first international*

conference on Machine learning [Internet]. New York, NY, USA: Association for Computing Machinery, 2004, 78. (ICML '04). https://doi.org/10.1145/1015330.1015435 (19 June 2021, date last accessed).

28. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR 2010;**9**:249–56. [Database]
29. Cox DR. The regression analysis of binary sequences. *J Royal Statis Soc Series B (Methodol)* 1958;**20**:215–32.
30. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 2020;**21**:6.
31. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta* 1975;**405**:442–51.
32. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 2019;**1**:206–15.
33. Ryo M, Angelov B, Mammola S *et al.* Explainable artificial intelligence enhances the ecological interpretability of black-box species distribution models. *Ecography* 2021;**44**:199–205.
34. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": explaining the predictions of any classifier. *arXiv:160204938* [cs, stat] [Internet]. 9 August 2016 [cited 10 June 2021], http://arxiv.org/abs/1602.04938 (19 June 2021, date last accessed).
35. Neubert P. *Superpixels and their Application for Visual Place Recognition in Changing Environments*. 2015. https://nbn-resolving.org/urn:nbn:de:bsz:ch1-qucosa-190241 (19 June 2021, date last accessed).
36. Vedaldi A, Soatto S. Quick shift and kernel methods for mode seeking. In: Forsyth D, Torr P, Zisserman A (eds), *Computer Vision—ECCV* [Internet]. Berlin, Heidelberg: Springer, 2008, 705–18. [cited 19 June 2021] http://link.springer.com/10.1007/978-3-540-88693-8_52 (19 June 2021, date last accessed).
37. Chawla NV, Bowyer KW, Hall LO *et al.* SMOTE: synthetic Minority Over-sampling Technique. *JAIR* 2002;**16**:321–57.
38. Nagy LG, Ohm RA, Kovács GM *et al.* Latent homology and convergent regulatory evolution underlies the repeated emergence of yeasts. *Nat Commun* 2014;**5**:4471.
39. Hibbett DS, Taylor JW. Fungal systematics: is a new age of enlightenment at hand? *Nat Rev Microbiol* 2013;**11**:129–33.
40. Hibbett D, Abarenkov K, Kõljalg U *et al.* Sequence-based classification and identification of Fungi. *Mycologia* 2016; **108**:1049–68.
41. Safonova A, Ghazaryan G, Stiller S *et al.* Ten deep learning techniques to address small data problems with remote sensing. *Int J Appl Earth Observ Geoinform* 2023;**125**:103569.
42. Tendle A, Hasan MR. A study of the generalizability of self-supervised representations. *Mach Learn Appl* 2021;**6**:100124.
43. Selvaraju RR, Cogswell M, Das A *et al.* Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int J Comput Vis* 2019;**128**:336–59.

Biology Methods and Protocols, 2024, 00, 1–11
https://doi.org/10.1093/biomethods/bpae063
Methods Article