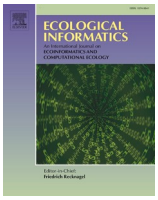
Ecological Informatics 99 (2026) 104098



Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



Deep learning for species identification in *Kirganelia* using digitized herbarium specimens

Eva Kristinawati Putri [a,b,c,*], Chiesa Serena [d], Lu Cao [d], Hans ter Steege [a,e], Erick Paulus [d,f], Peter C. van Welzen [a,b]

[a] *Naturalis Biodiversity Center, P.O. Box 9517, Leiden, 2300 RA, the Netherlands*
[b] *Institute of Biology Leiden, Leiden University, P.O. Box 9505, Leiden, 2300 RA, the Netherlands*
[c] *Universitas Negeri Surabaya, Jalan Lidah Wetan, Lidah Wetan, Surabaya, 60213, Indonesia*
[d] *Leiden Institute of Advanced Computer Science (LIACS), P.O. Box 9512, Leiden, 2300 RA, the Netherlands*
[e] *Quantitative Biodiversity Dynamics, Utrecht University, Budapestlaan 17, Utrecht, 3584 CD, the Netherlands*
[f] *Universitas Padjadjaran, Jalan Ir. Soekarno, Jatinangor, Sumedang, 45363, Indonesia*

ARTICLE INFO

Keywords:
Class imbalance
Deep learning
Herbarium specimens
Hierarchical evaluation
Kirganelia
Species identification

ABSTRACT

Accurate species identification of herbarium specimens is fundamental to biodiversity research but remains challenging in taxonomically complex plant groups. Here, we evaluated deep learning for species identification in *Kirganelia* (Phyllanthaceae) using 6475 cropped leaf images representing 30 phylogenomically validated species. Three convolutional neural network architectures were trained using transfer learning, and multiple strategies were compared to address severe class imbalance. Performance was evaluated using conventional image-level metrics and a hierarchical framework aggregating predictions at specimen and species levels. Model architectures showed different trade-offs between overall classification performance and recognition of under-represented species: EfficientNetV2-S provided strong aggregate performance, whereas MobileNetV3-L with weighted loss achieved the highest minority-class performance. Hierarchical evaluation showed that 19 of 24 sufficiently sampled species (79.2%) were well recognised, while the remaining five were moderately recognised. Five of six species represented by only one or two specimens were also correctly recognised for all available specimens, although these sparsely sampled taxa were evaluated descriptively. Grad-CAM analysis showed that predictions were primarily associated with leaf tissue rather than herbarium background. These findings support imbalance-aware training and hierarchical evaluation for accurate herbarium species recognition while highlighting the importance of class-specific performance, sampling representation, and expert verification.

1. Introduction

Herbarium collections provide indispensable reference material for taxonomy, biodiversity research, and conservation planning by preserving verifiable specimens and documenting species distributions across space and time (Bebber et al., 2010; Eckert et al., 2024; Guo et al., 2025; Marín-Rodulfo et al., 2024; Mayo et al., 2008). Yet, accurate identification remains challenging in taxonomically complex plant groups, particularly when diagnostic characters are subtle, inconsistent, or embedded in minute reproductive structures that are difficult to examine on actual and digitized herbarium sheets. These practical limitations highlight the need for automated, scalable, and non-destructive tools to support routine specimen identification.

Kirganelia Juss. presents a particularly suitable system for evaluating image-based species identification. Recent phylogenomic studies have clarified species relationships within the genus and established well-supported genus and species boundaries (Bouman et al., 2022). A recent phylogenomic revision of *Kirganelia* (Putri et al., 2026) analyzed 233 specimens and refined species boundaries using genome-scale evidence together with reproductive morphology, providing a stable taxonomic framework and reliable ground-truth labels for model training and evaluation. The study used Hyb-Seq, a target-enrichment and high-throughput sequencing approach widely applied in plant phylogenomics (Dodsworth et al., 2019; Weitemier et al., 2014), together with the Angiosperms353 universal probe set designed to recover 353 nuclear genes across flowering plants (Baker et al., 2022;

* Corresponding author at: Institute of Biology Leiden, Leiden University, P.O. Box 9505, 2300 RA Leiden, the Netherlands.
E-mail address: e.k.putri@biology.leidenuniv.nl (E.K. Putri).

https://doi.org/10.1016/j.ecoinf.2026.104098
Received 16 March 2026; Received in revised form 12 October 2026; Accepted 13 October 2026
Available online 14 October 2026
1574-9541/© 2026 Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Johnson et al., 2019). Despite this robust taxonomic framework, species identification remains difficult because the most informative diagnostic characters are minute reproductive structures, often less than 1 mm long, that require delicate dissection and may risk damaging herbarium specimens. Moreover, vegetative morphology is highly conserved across many species, providing few obvious characters for reliable visual identification. These characters make *Kirganelia* an ideal system for evaluating image-based species identification in a taxonomically challenging plant group.

The increasing digitization of herbarium collections, together with advances in computer vision, has created new opportunities for deep-learning-based species identification. Convolutional neural networks (CNNs) have consistently outperformed traditional feature-based approaches for classifying digitized herbarium specimens (Algemayel et al., 2025; Carranza-Rojas et al., 2017; Guo et al., 2025) and can achieve performance comparable to expert annotation in some applications (Ávila et al., 2025; Bonnet et al., 2018). Previous studies have also shown that focusing on localized plant structures, such as leaves or reproductive organs, improves classification by reducing the influence of non-informative regions, including labels, mounting paper, and overlapping plant parts (Belhumeur et al., 2008; Kajihara et al., 2025; Lorieul et al., 2019; Meineke et al., 2020). Accordingly, we used leaf-only images because leaves are the most consistently preserved organs in digitized herbarium specimens, whereas reproductive structures are frequently absent, damaged, or too small to examine reliably. Restricting the dataset to leaf images also reduced specimen-specific artifacts associated with complete herbarium sheets, an important consideration given the limited availability of phylogenetically validated specimens for several species. Nevertheless, herbarium datasets are inherently imbalanced because rare and geographically restricted taxa are often represented by only a few specimens, potentially biasing CNN predictions toward common species and reducing generalization performance (Buda et al., 2018; Lorieul et al., 2019; Meineke et al., 2020). These challenges highlight the importance of evaluating deep-learning models under realistic taxonomic conditions characterized by subtle morphological variation and highly uneven species representation.

This study represents the first application of deep learning to species identification in *Kirganelia*, a taxonomically challenging lineage with species boundaries independently supported by phylogenomic and reproductive-morphological evidence. Rather than proposing a new deep-learning architecture, we present a biologically informed framework for evaluating deep-learning performance using digitized herbarium specimens. Specifically, we (1) evaluate imbalance-aware training strategies under conditions of severe specimen scarcity and unequal species representation, (2) implement a hierarchical evaluation framework integrating image-, specimen-, and species-level assessments to provide a biologically meaningful assessment of model performance, and (3) relate classification performance to phylogenomic relationships and morphological differentiation among species. More broadly, this work demonstrates how phylogenomically validated training data can support routine identification of digitized herbarium specimens while reducing the need for destructive examination of minute reproductive structures. This biologically informed framework may also be applicable to other taxonomically complex plant groups.

2. Methods

2.1. Data acquisition

The dataset used in this study was based on the phylogenomic sampling of *Kirganelia* presented in Putri et al. (2026). For each specimen included in the phylogenomic revision, the corresponding herbarium sheet image was retrieved primarily through the BioPortal platform of Naturalis Biodiversity Center, the Netherlands (https://bioportal.naturalis.nl/) or, when necessary, from other participating herbaria. By restricting the dataset to specimens included in the phylogenomic framework, we ensure that each herbarium image is associated with a taxonomically validated species assignment supported by molecular evidence. In rare cases where a sequenced specimen lacked a digitized image or the available image was insufficient for leaf-based analysis, a single well-identified herbarium specimen of the same species, verified through consistent morphology and geographic origin, was included to maintain dataset completeness while remaining aligned with the taxonomic framework.

The final dataset (Supplementary 1) comprised 314 herbarium specimens representing 30 *Kirganelia* species, each represented by a full herbarium sheet image prior to processing. Fig. 1 summarizes the number of specimens per species and illustrates the uneven distribution of samples across the dataset.

2.2. Data pre-processing

All herbarium sheet images were pre-processed prior to model training. Pre-processing consisted of (i) resizing images to a fixed resolution, (ii) automated cropping to isolate leaf regions from the full sheet, and (iii) manual removal of patches that did not contain target material (e.g., labels, envelopes, or non-leaf structures). These steps standardized image dimensions and minimized non-informative content in the training data. The preprocessing pipeline is illustrated in Fig. 2, while representative herbarium sheets illustrating the challenges of the dataset are presented in Fig. 3.

Resizing standardized the physical scale of herbarium sheet images obtained from multiple herbaria using different scanning resolutions. A reference ruler based on Naturalis Biodiversity Center images at 100% magnification was used to compare scale bars, and images from other collections were proportionally rescaled when necessary to ensure a consistent physical scale across the dataset.

Leaf regions were automatically cropped using a custom Python script that combines color, intensity, edge, and texture information. Color segmentation used empirically determined thresholds in the HSV (H: 20–80, S: 10–100, V: 40–200) and CIELAB color spaces. Additional masks were generated from grayscale intensity thresholds, Laplacian texture responses (Marr and Hildreth, 1980), Sobel gradients (Sobel, 1976), and Canny edge detection (Canny, 1986). These masks were combined using majority voting and refined by background removal and morphological operations to produce a continuous leaf mask. A non-overlapping sliding window (224 × 224 pixels) was then applied, retaining patches containing at least 60% leaf pixels. This patch size corresponds to the standard input resolution of ImageNet-pretrained CNN architectures (He et al., 2016; Tan and Le, 2021). The resulting dataset comprised 12,946 cropped leaf images representing 30 *Kirganelia* species. The preprocessing script is publicly available in the project repository (see Data Availability).

2.3. Class imbalance handling

The resulting dataset exhibited pronounced class imbalance, with class sizes ranging from 8 to more than 1000 images (mean ≈ 431 images per species; Fig. 4). Automatic leaf cropping further amplified differences in specimen representation because each herbarium sheet yielded a variable number of leaf patches depending on leaf size, position, and image quality. Such uneven class representation can bias CNN models toward majority classes and reduce recognition performance for underrepresented species.

To reduce the impact of class imbalance during training, three complementary strategies were applied. First, class-weighted cross-entropy was used, with weights derived from both inverse-frequency and effective-number formulations to account for differences in class sizes (Cui et al., 2019; King and Zeng, 2001). Second, focal loss was employed to emphasize harder training examples and prevent easy cases, typically belonging to majority classes, from dominating the loss (Lin et al., 2017). Third, balanced sampling was implemented by drawing

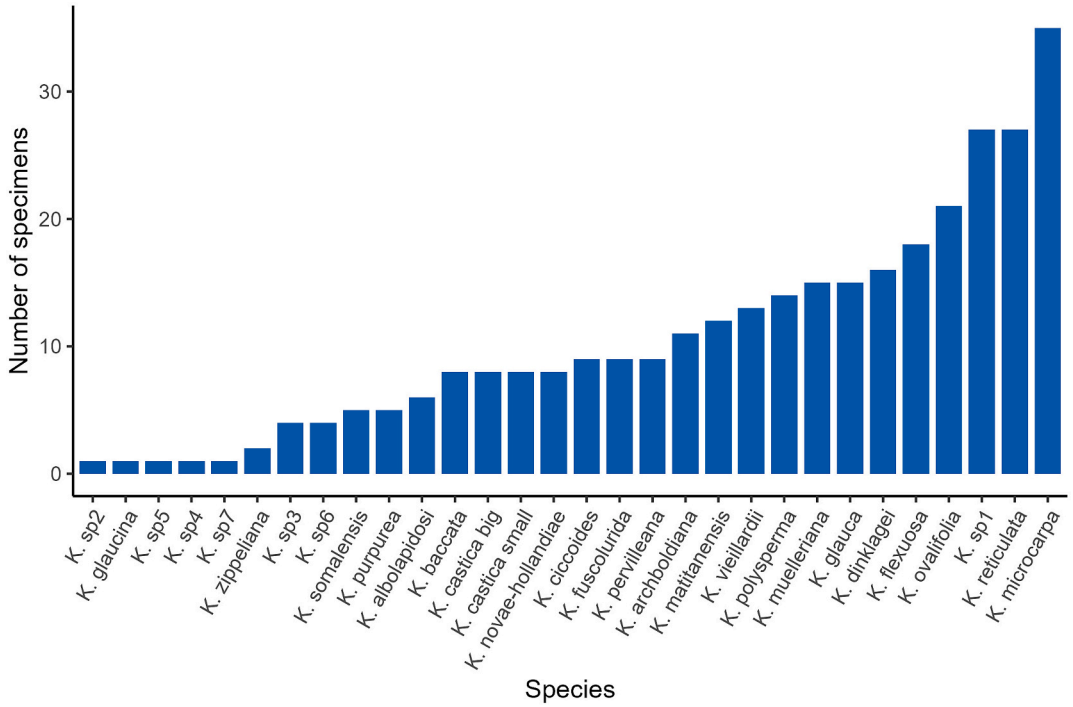


Fig. 1. Distribution of the number of herbarium specimens per *Kirganelia* species used in this study. Bars represent the number of specimens available for each species, highlighting substantial class imbalance across the dataset, with several species represented by only a few specimens and others by more than 10 specimens.

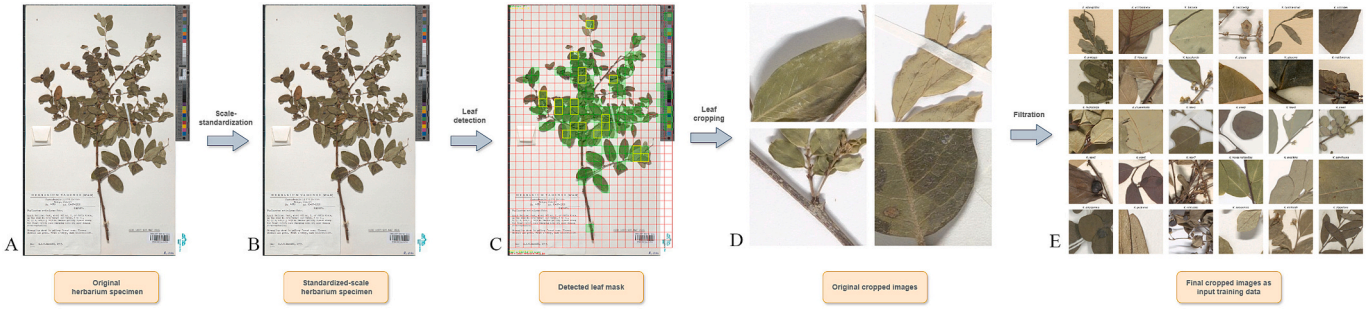


Fig. 2. Pre-processing pipeline for herbarium images. Herbarium sheets were standardized to a common scale, leaf regions were detected and cropped, and unsuitable patches were filtered. The resulting leaf images were used as input for CNN-based species classification.

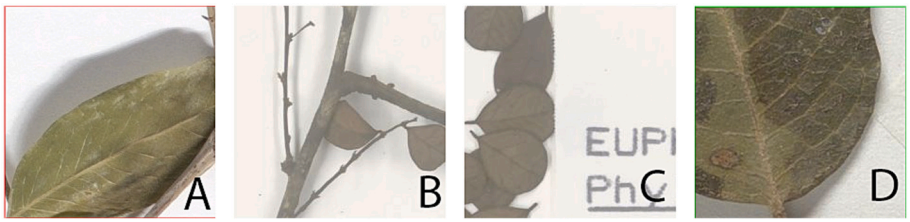


Fig. 3. Examples of image patches removed during filtering. Patches containing strong shadows (A), excessive stem (B), or labels © were excluded. Patches with clear leaf morphology (D) were retained for model training.

training images with probabilities inversely proportional to species frequency, ensuring more frequent exposure to minority classes (Buda et al., 2018). These strategies were applied individually and in combination during model training.

2.4. Evaluation metrics

This study uses multiple evaluation metrics to account for the strong class imbalance in the dataset, as relying solely on accuracy can be misleading (Hastie et al., 2009). Overall accuracy is reported as a general performance indicator, but because it is biased toward majority classes, balanced accuracy is also included and is defined as the average per-class recall (Brodersen et al., 2010). Macro-F1 averages the F1 score equally across classes, giving identical weight to each class regardless of its size, whereas Weighted F1 weights each class according to its support (Sokolova and Lapalme, 2009). Minority F1 (computed as the mean F1

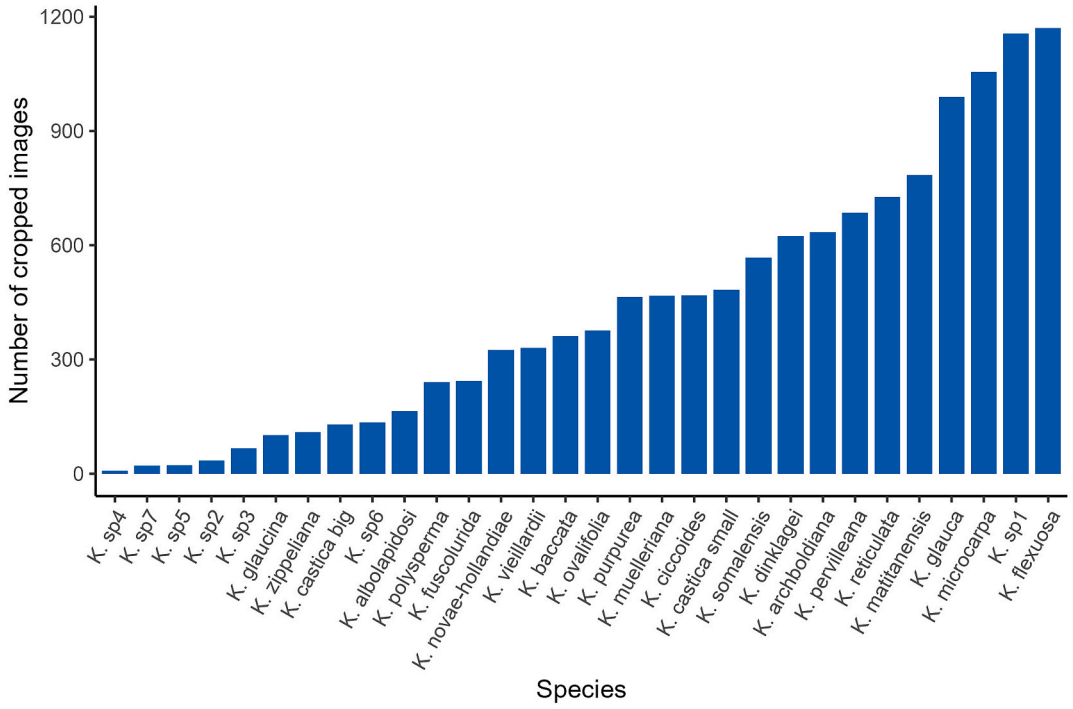


Fig. 4. Number of cropped leaf images obtained per *Kirganelia* species. Following automated sliding-window extraction and filtering based on leaf coverage, the number of retained image patches varies widely among species, reflecting pronounced class imbalance in the training dataset.

score over underrepresented species) specifically evaluates model performance on rare classes. In addition, Cohen's kappa is used to measure label agreement while adjusting for chance effects (Cohen, 1960). Mathematical definitions of all evaluation metrics are provided in Table 1. Finally, confusion matrices were generated to visualize species-level prediction errors and provide a detailed breakdown of misclassifications (Hastie et al., 2009).

2.5. Experimental design

The dataset was divided into training, validation, and test subsets using stratified sampling to preserve species proportions. Hyperparameter tuning was performed using five-fold stratified cross-validation, whereas the independent test set was reserved exclusively for final performance evaluation. Stratification was performed at the image-patch level to preserve species proportions across the training, validation, and test sets. Patches originating from the same herbarium specimen were not grouped during data partitioning and could therefore be assigned to different data splits.

The initial experiment compared six class imbalance handling strategies (Table 2), including a baseline without imbalance correction, class-weighted loss, balanced sampling, focal loss, focal loss combined with class weighting, and a fully combined approach integrating focal loss, class weighting, and balanced sampling. Each strategy was evaluated after 5, 10, 15, and 20 training epochs using three CNN architectures: ResNet50, EfficientNetV2-Small, and MobileNetV3-Large. These models were selected to represent different trade-offs between model capacity and computational efficiency. The best-performing configuration was subsequently used for hyperparameter tuning and final model training.

Following the initial experiment, hyperparameter tuning was performed for each architecture under two configurations: a baseline without imbalance handling and the best-performing imbalance strategy identified previously. Up to 50 hyperparameter configurations were evaluated using random search supplemented with several manually selected combinations (Bergstra and Bengio, 2012), with the explored search space summarized in Table 3. Early stopping with a patience of three epochs was applied within each fold. The final hyperparameter configuration was selected based on cross-validation performance and used for final model training, while the independent test set was

Table 1
Evaluation metrics used to assess classification performance.

Metric	Formula
Accuracy	$\text{Accuracy} = (\sum_{i=1}^{C} \text{TP}_i) / N$
Recall	$\text{Recall}_i = \text{TP}_i / (\text{TP}_i + \text{FN}_i)$
Precision	$\text{Precision}_i = \text{TP}_i / (\text{TP}_i + \text{FP}_i)$
F1-score	$\text{F1}_i = (2 \times \text{Precision}_i \times \text{Recall}_i) / (\text{Precision}_i + \text{Recall}_i)$
Balanced Accuracy	$\text{BalancedAccuracy} = (1 / C) \times \sum_{i=1}^{C} \text{Recall}_i$
Macro F1	$\text{MacroF1} = (1 / C) \times \sum_{i=1}^{C} \text{F1}_i$
Weighted F1	$\text{WeightedF1} = \sum_{i=1}^{C} (n_i / N) \times \text{F1}_i$
Minority F1	$\text{MinorityF1} = (1 / M) \times \sum_{i \in \text{Minority}} \text{F1}_i$
Cohen's κ	$\kappa = (p_o - p_e) / (1 - p_e)$

Accuracy, Recall, Precision, F1-score, Balanced Accuracy, Macro F1, Weighted F1, Minority F1, and Cohen's κ are reported together with their corresponding formulas. Abbreviations: TP_i, FP_i, and FN_i – true positives, false positives, and false negatives for class i; C – total number of classes; M – number of minority classes; n_i – number of samples in class i; N – total number of samples; p_o – observed agreement; p_e – expected agreement by chance.

Table 2
Loss functions and sampling strategies evaluated in this study.

Loss Method	Class Weights	Focal Loss	Balanced Sampling
Baseline (Cross-entropy loss)	No	No	No
Weighted Loss	Yes	No	No
Focal Loss	No	Yes	No
Focal + Weighted Loss	Yes	Yes	No
Balanced Sampling	No	No	Yes
All Combined	Yes	Yes	Yes

The table summarizes the combinations of techniques used to address class imbalance during model training. "Class Weights" indicates whether class weighting was applied in the loss function, "Focal Loss" indicates the use of focal loss, and "Balanced Sampling" denotes whether balanced sampling was applied during batch generation.

Table 3
Hyperparameters and search ranges used during model optimization.

Hyperparameter	Search Type	Range / Options
Learning Rate	Continuous (log scale)	1e-5 – 1e-3
Batch Size	Discrete	{16, 32, 64}
Optimizer	Categorical	{Adam, AdamW}
Weight Decay	Continuous	0 – 1e-3
Dropout Rate	Continuous	0.0–0.5
Learning Rate Scheduler	Categorical	{None, CosineAnnealing, StepLR}
Label Smoothing	Continuous	0.0–0.2
Gradient Clipping	Continuous	0–1.0
Warm-up Epochs	Discrete	{0, 3, 5}

The table summarizes the hyperparameters included in the tuning procedure together with their corresponding search types and explored ranges or options. Continuous parameters were sampled within the specified ranges, while discrete and categorical parameters were selected from predefined sets.

reserved solely for final evaluation.

Using the selected hyperparameters, class imbalance strategy, and training duration, each architecture was retrained and evaluated using the cross-validation framework described above. In each cross-validation iteration, models were trained with the fixed configuration and evaluated on the held-out test set. Performance metrics were averaged across cross-validation runs to obtain the final performance estimates for each configuration.

2.6. Specimen-level and species-level recognition analysis

Because each herbarium specimen generated multiple cropped leaf images, model performance was additionally evaluated at the specimen and species levels to provide a biologically meaningful interpretation. Herbarium accession numbers extracted from image filenames (Supplementary 1) were used to group cropped images originating from the same specimen. Specimen identifiers were parsed automatically by removing filename extensions and crop suffixes (e.g., _0003) and extracting herbarium codes and accession numbers using regular expressions.

For each specimen, predictions from all cropped images were aggregated hierarchically across the five cross-validation folds. Within each fold, accuracy was calculated as the proportion of correctly classified cropped images, and specimen-level accuracy was obtained by averaging these values across folds. A specimen was considered recognised when its accuracy was at least 0.80.

Species-level recognition performance was quantified as the proportion of recognised specimens within each species (specimen recognition rate). Species represented by fewer than three specimens were considered insufficiently sampled for reliable inference and are reported descriptively only. Species represented by three or more specimens were classified as well recognised (recognition rate ≥ 0.80), moderately recognised (0.50–0.79), or poorly recognised (< 0.50).

For moderately recognised species, prediction confidence was further analyzed at the specimen level. Mean confidence was first calculated across cropped images within each fold and then averaged across folds. Confidence categories (HIGH, MEDIUM, and LOW) were assigned using the modal category for each specimen. At the species level, overall mean confidence and the proportions of specimens in each confidence category were calculated.

Misidentification patterns among moderately recognised species were examined using a specimen-level confusion matrix. Predicted labels were first aggregated within each fold by majority voting and subsequently combined across folds by majority voting to obtain final specimen-level predictions. These were compared with the corresponding true species labels to identify recurrent alternative species assignments, and the resulting confusion matrix was visualized as a heatmap. All analyses and visualizations were performed using custom Python scripts executed in Google Colab (Python 3.12) with pandas (McKinney, 2010), NumPy (Harris et al., 2020), scikit-learn (Pedregosa et al., 2011), and matplotlib (Hunter, 2007).

2.7. Characterization of morphological features used by the CNN

To characterize the image regions contributing to species recognition, Gradient-weighted Class Activation Mapping (Grad-CAM) was applied to the best-performing CNN model using all cropped herbarium images. For qualitative assessment, up to ten test images were randomly sampled from each *Kirganelia* species using a fixed random seed (42), yielding a representative subset of 298 images. One species contributed fewer than ten images due to limited test samples. The corresponding Grad-CAM visualizations were manually examined to determine whether the dominant activation was concentrated on biologically relevant leaf tissue or on non-biological image features (e.g., mounting paper, tape, thread, or lighting artifacts), which could indicate shortcut learning (Selvaraju et al., 2017).

3. Results

3.1. Effects of class imbalance handling and training duration

The comparison of class imbalance handling strategies revealed distinct performance patterns across model architectures. Across all models, improvements in minority-class recognition were often accompanied by trade-offs in overall classification performance or training stability.

Different imbalance-handling strategies were optimal for different architectures. EfficientNetV2-S showed relatively stable performance across strategies. Although focal loss improved consistency for some underrepresented classes, overall Balanced Accuracy and Macro-F1 remained comparable to the baseline. MobileNetV3-L achieved its best performance with weighted loss, showing the strongest minority-class recognition together with high Balanced Accuracy and overall classification performance. ResNet50 showed moderate improvement under the combined focal and weighted loss configuration, although performance varied more across strategies than in the other architectures.

Based on the combined evaluation of overall classification performance, minority-class recognition, training stability, Macro-F1, Balanced Accuracy, and Cohen's kappa, focal loss was selected for EfficientNetV2-S, weighted loss for MobileNetV3-L, and focal + weighted loss for ResNet50 for subsequent experiments. Performance gains beyond 20 epochs were marginal; therefore, 20 epochs were adopted for subsequent experiments. Figs. 5–7 show the performance distributions for all evaluated strategies.

3.2. Hyperparameters and minority-class performance

Figs. 8–10 show the linear relationships between individual hyperparameters and Minority F1 performance across the three evaluated CNN architectures.

Most hyperparameters exhibited weak or inconsistent linear relationships with Minority F1. In EfficientNetV2-S, regression slopes were generally shallow across the evaluated hyperparameter ranges, indicating limited linear associations with Minority F1. In MobileNetV3-L, learning rate showed the strongest negative linear relationship with Minority F1, whereas weight decay and label smoothing exhibited weak positive relationships. In ResNet50, most hyperparameters also showed weak linear relationships, with only batch size and dropout rate displaying slight positive trends.

Comparisons among loss configurations showed that weighted loss achieved the highest Minority F1 values in MobileNetV3-L, whereas focal loss produced more consistent Minority F1 performance in EfficientNetV2-S. ResNet50 exhibited greater variability across the evaluated training settings than the other two architectures. Despite

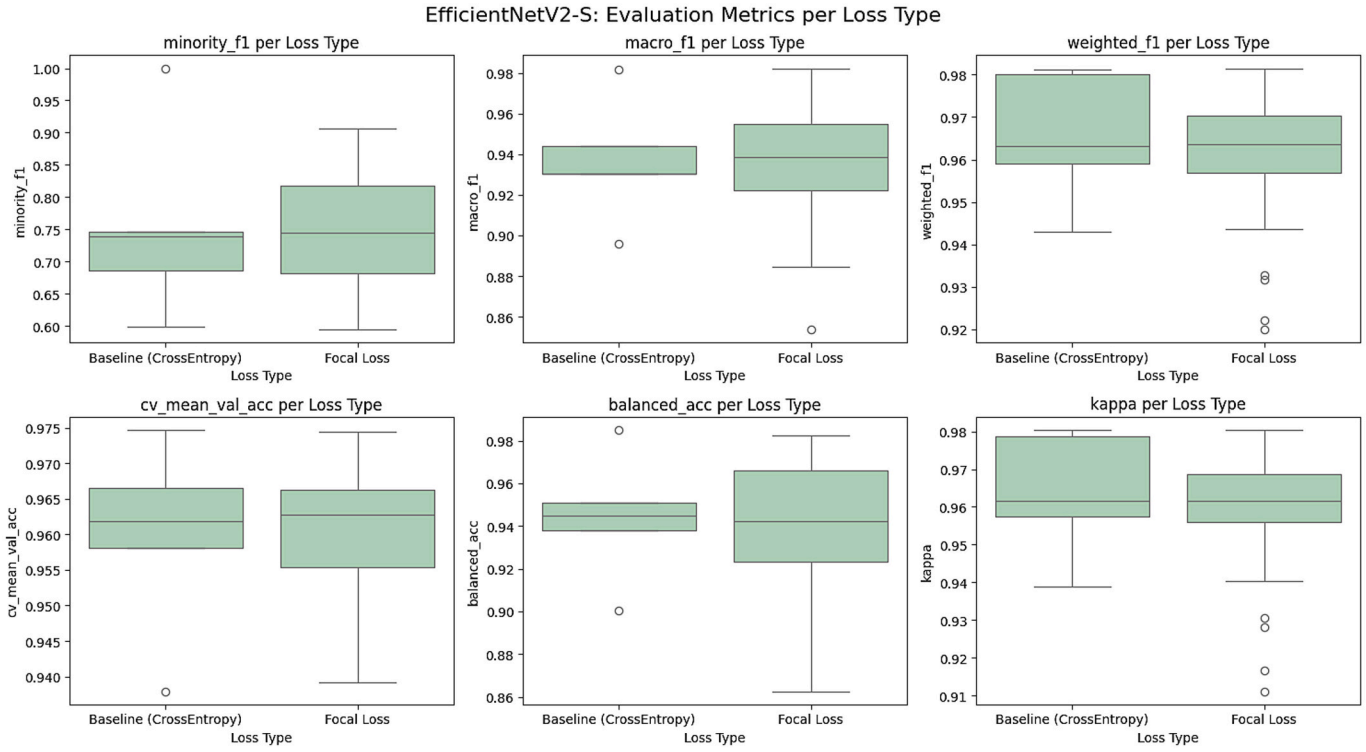


Fig. 5. Distribution of evaluation metrics across loss types for EfficientNetV2-S. Focal loss improves minority F1 while maintaining other metrics.

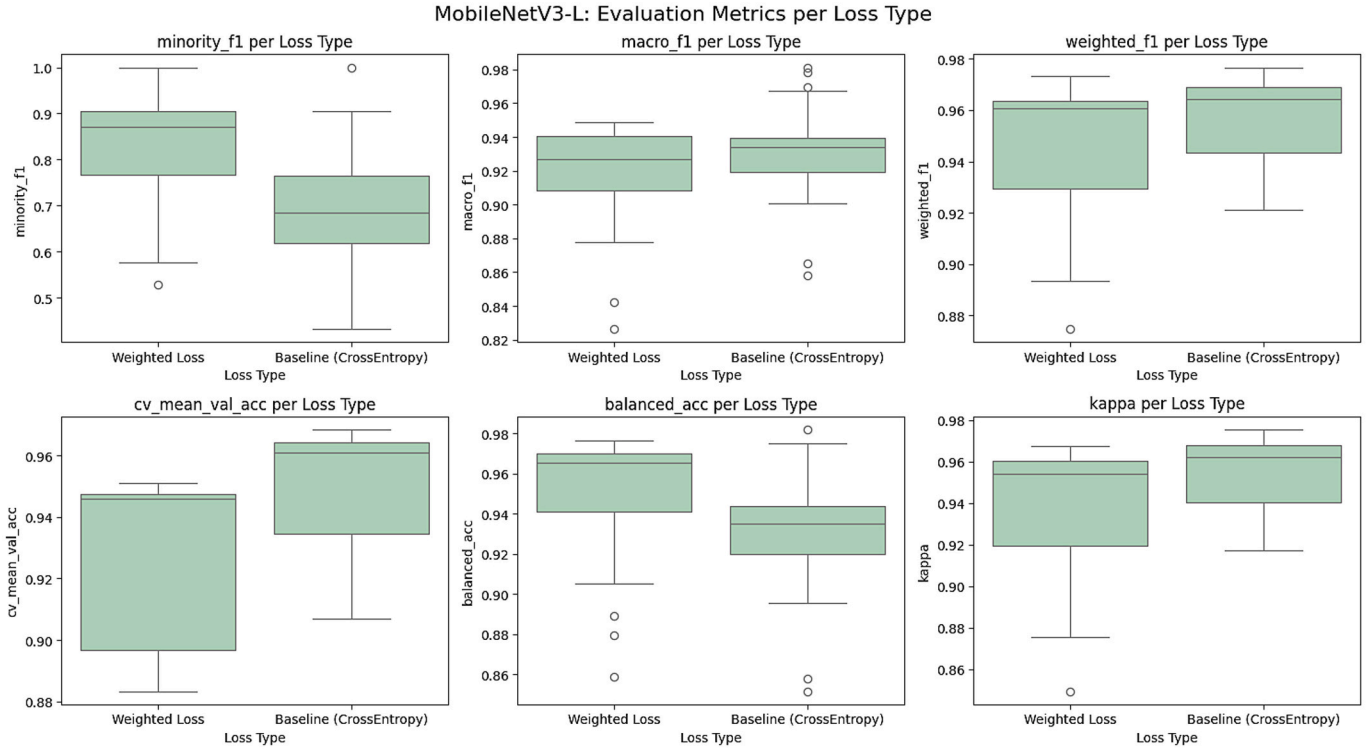


Fig. 6. Distribution of evaluation metrics across loss types for MobileNetV3-L. Weighted loss improves minority F1 significantly while maintaining stable macro-F1, accuracy, and kappa.

these differences, improvements in Minority F1 were generally not accompanied by substantial reductions in Macro-F1, Balanced Accuracy, or Cohen's κ.

3.3. Final model performance comparison

Final model performance was evaluated using the selected imbalance-handling strategies and optimized hyperparameters for each

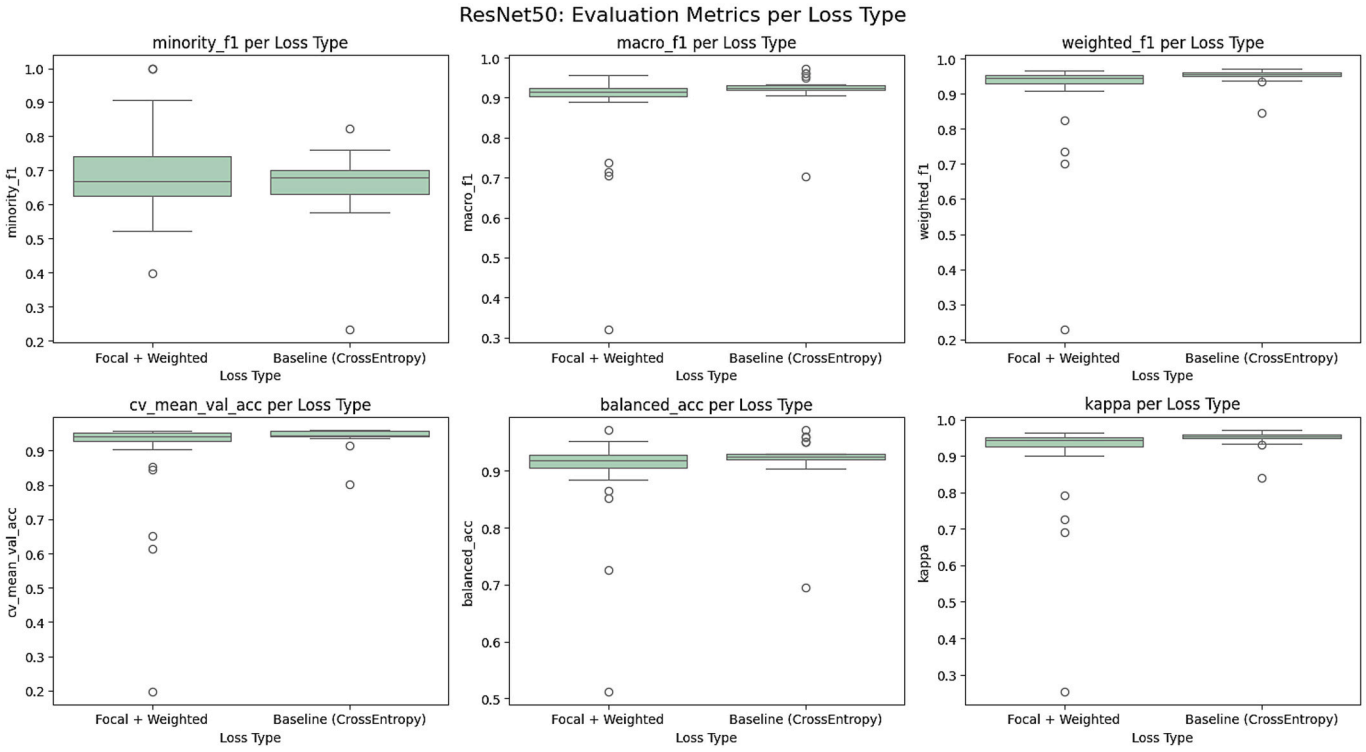


Fig. 7. Distribution of evaluation metrics across loss types for ResNet50. Applying focal loss combined with weighted loss improves minority F1 score while maintaining stable other metrics.

architecture. Performance metrics are summarized in Table 4.

The final evaluation revealed trade-offs between overall classification performance and minority-class recognition across architectures. EfficientNetV2-S achieved the highest Accuracy, Weighted F1, and Cohen's κ among the evaluated models. However, its Minority F1 showed greater variability across cross-validation folds than the other architectures.

MobileNetV3-L with weighted loss achieved the highest Minority F1 and Balanced Accuracy while maintaining competitive performance for the remaining evaluation metrics. Performance variability across folds was also comparatively low. In contrast, ResNet50 showed greater variability and smaller improvements under the combined focal and weighted loss strategy.

3.4. *Specimen-level and species-level recognition performance*

Specimen-level aggregation was used to evaluate recognition consistency across biological individuals rather than individual cropped images (Supplementary 2). Of the 30 *Kirganelia* species included in the dataset, specimen representation per species ranged from 1 to 21, with most species represented by fewer than 10 specimens. Among these, 24 species were represented by at least three specimens and were included in categorical recognition analyses, whereas six species were represented by only one or two specimens and were therefore reported descriptively only.

Among the six insufficiently sampled species, five were well recognised for all available specimens (Fig. 11), whereas one species (*sp4*), classified as poorly recognised. For *sp4*, two cropped images were predicted as *K. albolapidosi* in the first-ranked prediction. However, the correct species identity (*sp4*) appeared consistently as the second-ranked prediction, with confidence values of 20.9% and 37.0%.

Among the 24 sufficiently sampled species, 19 species (79.2%) were classified as well recognised, with specimen recognition rates of 0.80 or higher (Fig. 11, Supplementary 3), whereas the remaining five species (20.8%) were classified as moderately recognised with recognition rates ranging from 0.67 to 0.79, representing a relatively narrow interval within the moderate-recognition category. Despite variation among species, all moderately recognised taxa were correctly identified for the majority of their specimens.

Confidence-category composition among moderately recognised species is shown in Fig. 12. Although all five species showed moderate specimen recognition rates, substantial differences were observed in their confidence profiles. *Kirganelia castica-big* was dominated by HIGH-confidence specimens, whereas *K. glauca* consisted entirely of LOW-confidence specimens. *Kirganelia fuscolurida* and *K. ciccoides* showed mixed distributions of HIGH- and MEDIUM-confidence specimens, while *K. reticulata* showed a combination of MEDIUM- and LOW-confidence specimens.

Specimen-level misidentification patterns for moderately recognised species are shown in Fig. 13. Correct identifications were more frequent than misidentifications across all moderately recognised species, with diagonal values dominating the confusion matrix. Misidentification events were limited in number and involved only a small set of alternative species, including *K. microcarpa* and the species labelled *sp4* and *sp6*, rather than widespread confusion across multiple species.

3.5. *Morphological features used by the CNN*

Gradient-weighted Class Activation Mapping (Grad-CAM) was used to visualize the image regions contributing to species recognition by the best-performing CNN model. A qualitative assessment of 298 randomly sampled test images representing all 30 *Kirganelia* species showed that the dominant activation was consistently centered on leaf tissue (Fig. 14). This leaf-dominated activation pattern was observed both in cropped images containing only leaves and in images that also included stems, twigs, flowers, or fruits. When additional plant structures were present, they were either not encompassed by the activation map (Fig. 14A, B) or were only partially covered by the diffuse leaf-centered activation, without forming distinct activation hotspots (Fig. 14C). Furthermore, the activation generally covered broad regions of the leaf

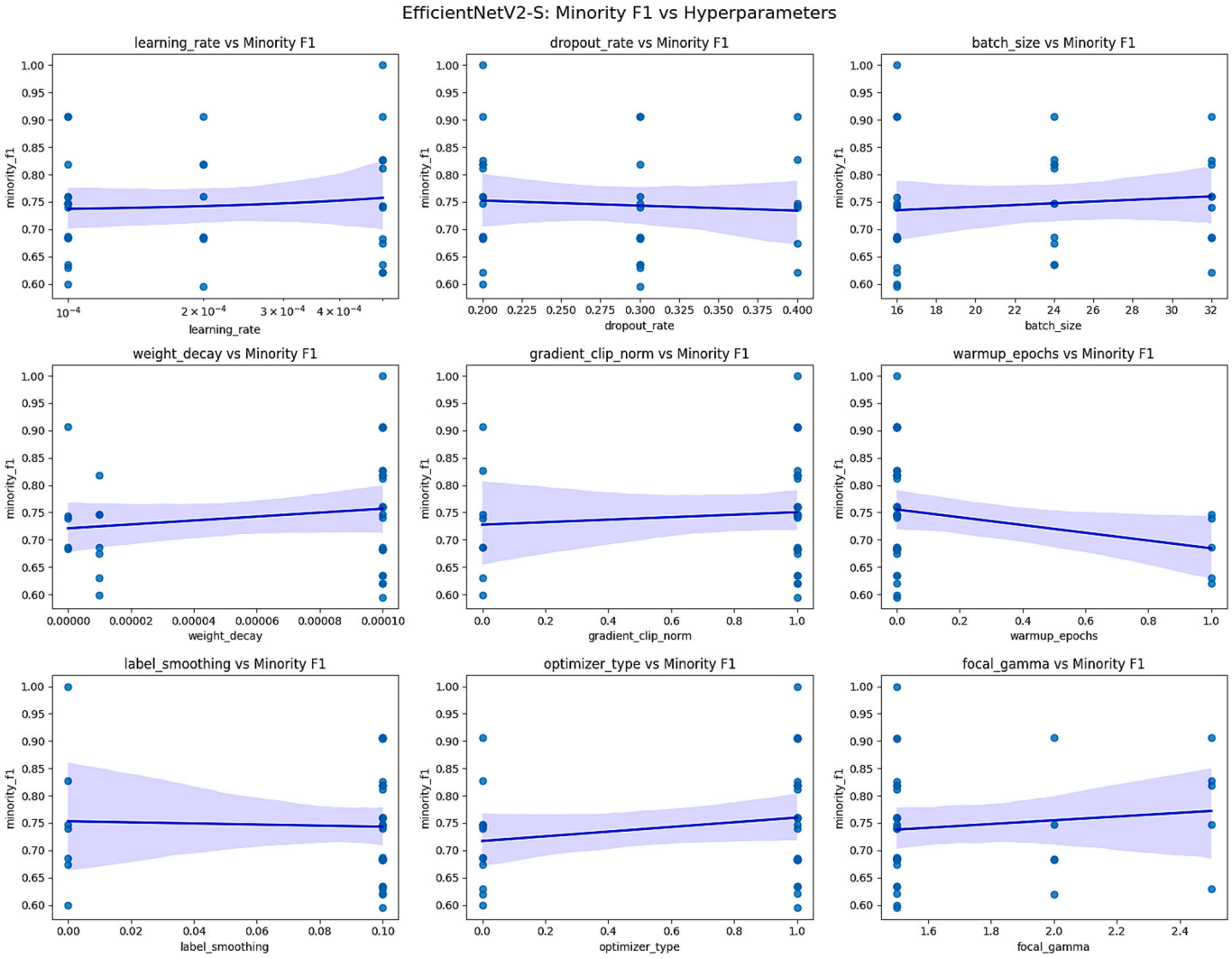


Fig. 8. Linear relationships between individual hyperparameters and Minority F1 for EfficientNetV2-S. Most hyperparameters exhibited weak linear relationships with Minority F1.

rather than being localized to specific morphological features such as the leaf margin, apex, base, or venation.

Dominant activation on non-biological image features was not observed in the sampled images. An exception was observed in several incorrectly classified images of *Kirganelia sp4*, in which activation occasionally extended to mounting tape or thread.

4. Discussion

This study demonstrates that deep learning models can achieve high species identification performance from leaf-only herbarium images, even for a taxonomically challenging group such as *Kirganelia*. However, classification performance varied substantially among species, particularly under severe class imbalance. These findings highlight the trade-offs between maximizing overall classification accuracy and achieving reliable identification of underrepresented species, while also demonstrating both the potential and limitations of image-based species identification for taxonomically complex plant groups.

4.1. Model architecture and responses to class imbalance

The results indicate that the effectiveness of class imbalance handling depends on the underlying CNN architecture rather than representing a universally applicable solution. Although all three architectures achieved high overall classification performance, their responses to imbalance correction differed considerably, particularly for minority-class recognition. This finding suggests that selecting an appropriate imbalance handling strategy is as important as selecting the model architecture itself when developing image-based species identification systems.

MobileNetV3-L responded most consistently to weighted loss, showing substantial improvements in minority-class performance while maintaining competitive overall accuracy. By assigning greater importance to misclassified minority-class samples during optimization, weighted loss encourages the model to allocate more attention to underrepresented classes. Previous studies have similarly emphasized that the effectiveness of class imbalance handling depends on the characteristics of the learning problem and that no single strategy is universally optimal (Johnson and Khoshgoftaar, 2019).

EfficientNetV2-S, in contrast, achieved strong overall performance even without imbalance correction but showed greater variability in minority-class metrics under baseline training. Applying focal loss reduced this variability and produced more consistent minority-class performance across cross-validation folds. Because focal loss down-weights the loss contribution of well-classified examples and focuses training on harder examples (Lin et al., 2017), it appeared to provide a more balanced optimization strategy for this architecture under the class distribution examined in the present study.

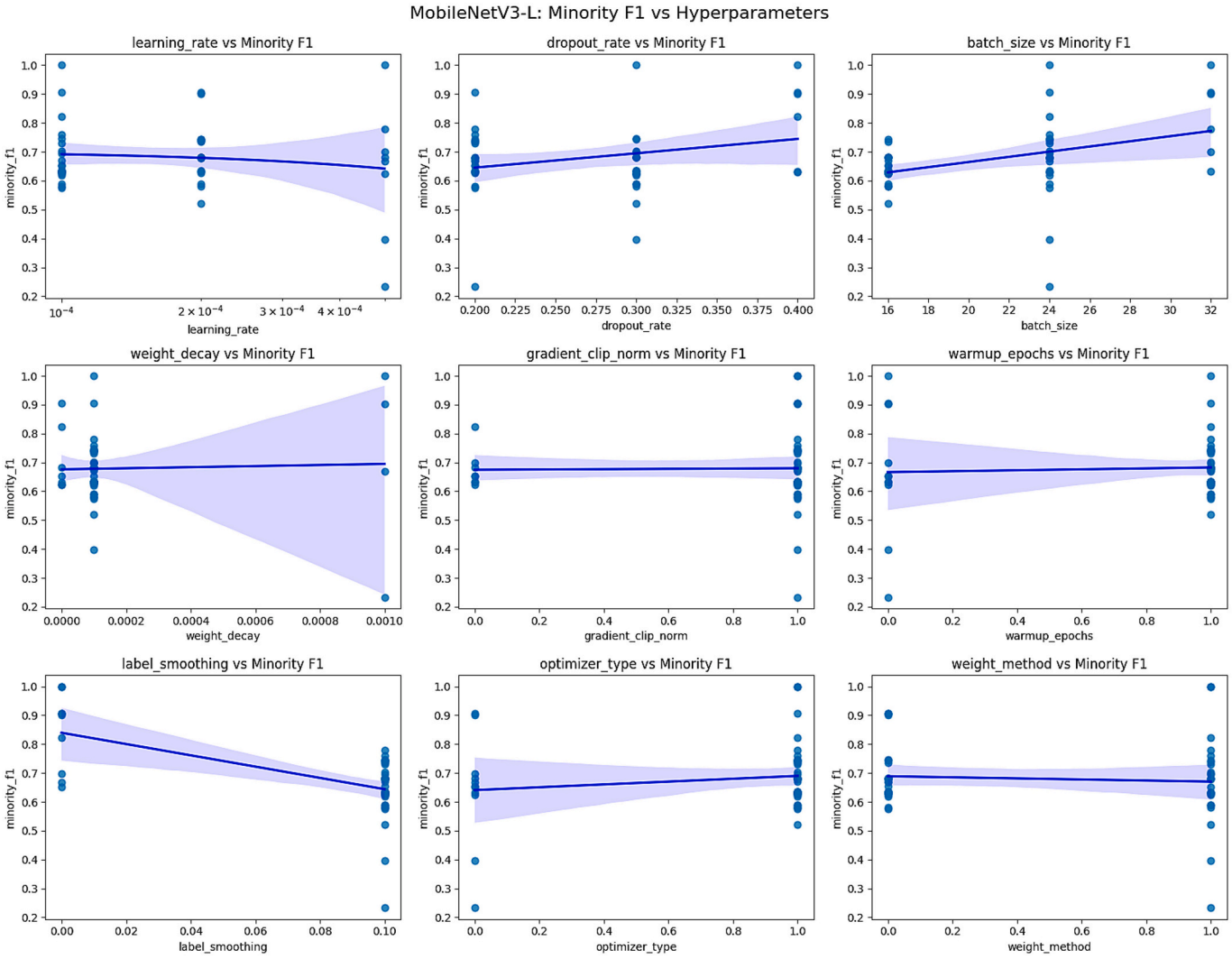


Fig. 9. Linear relationships between individual hyperparameters and Minority F1 for MobileNetV3-L. Learning rate exhibited the strongest negative linear relationship with Minority F1, whereas weight decay and label smoothing showed weak positive relationships.

ResNet50 showed comparatively limited improvement when multiple imbalance correction strategies were combined. Although focal loss and weighted loss individually improved minority-class performance during earlier experiments, their combination did not produce additional gains during the final model evaluation. These findings indicate that combining multiple imbalance handling methods does not necessarily yield additive benefits and that the effectiveness of such strategies depends on the characteristics of the underlying CNN architecture.

4.2. *Trade-offs between overall and minority-class performance*

A key finding of this study is the persistent trade-off between maximizing overall classification performance and improving recognition of underrepresented species. EfficientNetV2-S achieved the highest overall Accuracy, Weighted F1, and Cohen's κ, but its Minority F1 was lower and considerably more variable across folds than that of MobileNetV3-L with weighted loss. In contrast, MobileNetV3-L with weighted loss achieved the highest Minority F1 and Balanced Accuracy, while showing only a modest reduction in overall classification performance. These results demonstrate that the configuration providing the strongest aggregate performance is not necessarily the one providing the most reliable recognition of underrepresented species.

The practical importance of this trade-off depends on the intended application of the identification system. In large-scale herbarium digitization or routine specimen identification, maximizing overall classification performance may be appropriate when most specimens belong to well-represented taxa. In contrast, biodiversity assessment and taxonomic studies may place greater value on correctly recognizing poorly represented taxa, for which minority-class performance becomes particularly important. Under these circumstances, a modest reduction in aggregate performance may be acceptable if it improves recognition of underrepresented species. Consequently, model evaluation should consider both overall and class-specific metrics, as aggregate accuracy alone may provide an incomplete assessment of performance on imbalanced datasets (Haixiang et al., 2017).

For curatorial applications, these contrasting model behaviours also suggest a potential two-stage recognition strategy rather than relying on a single classifier for all specimens. A general classifier optimized for overall recognition could first be used for routine identification, with high-confidence predictions retained for curatorial assessment. Low-confidence or ambiguous specimens could then be evaluated using a second classifier optimized for minority-class recognition. In the present experiments, EfficientNetV2-S provided the strongest aggregate performance, whereas MobileNetV3-L with weighted loss provided the strongest minority-class performance, illustrating the potential complementarity of these approaches. Such a workflow could be particularly useful in collections in which a small number of taxa are represented by many specimens while others remain data deficient.

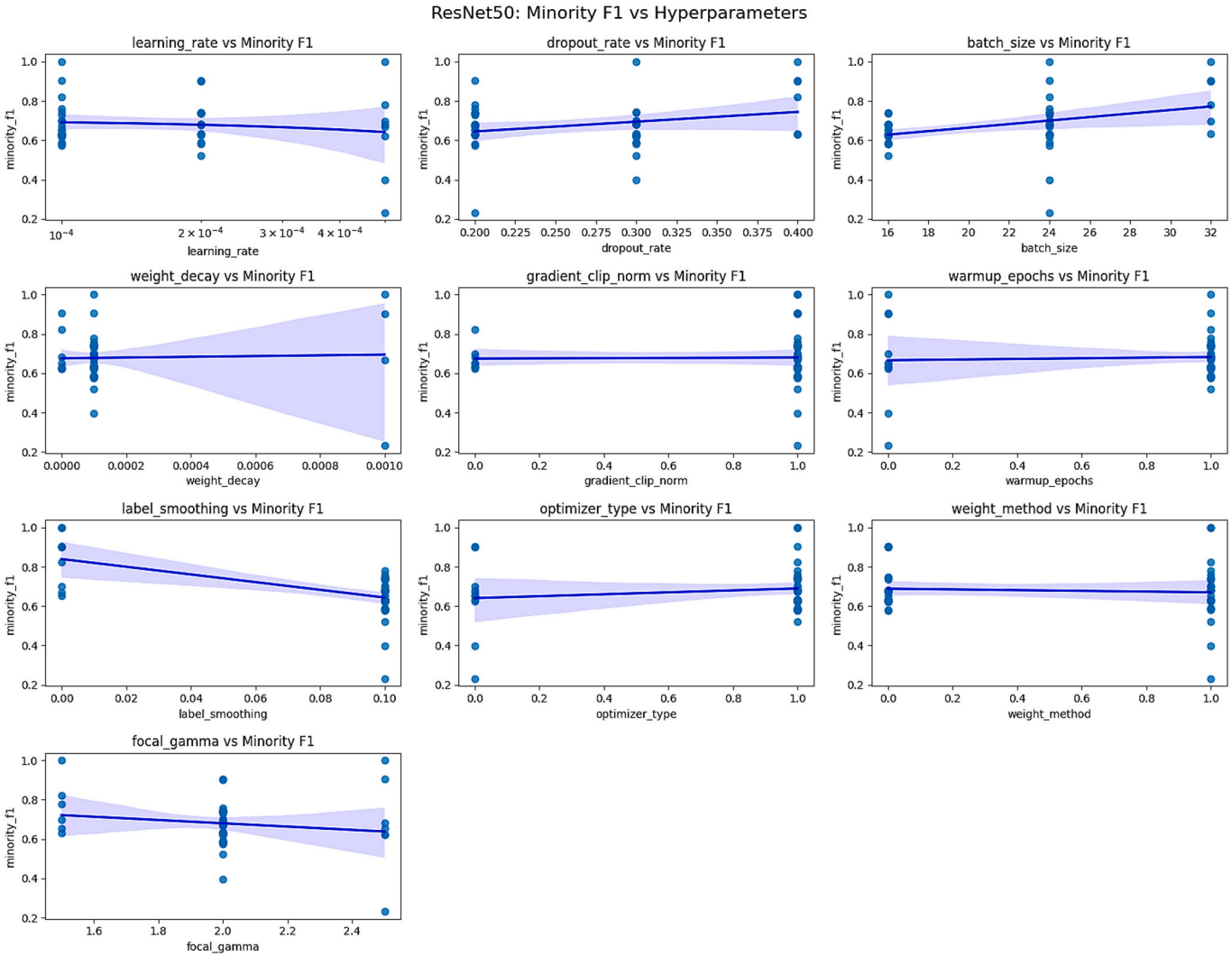


Fig. 10. Linear relationships between individual hyperparameters and Minority F1 for ResNet50. Most hyperparameters exhibited weak or inconsistent linear relationships with Minority F1, whereas batch size and dropout rate showed slight positive linear relationships.

Table 4
Performance comparison of different convolutional neural network models and imbalance-handling strategies.

Model	Method	Accuracy	Balanced Accuracy	Macro F1	Weighted F1	Minority F1	Kappa
EfficientNetV2-S	Baseline	**0.9750 ± 0.0013**	0.9573 ± 0.0176	0.9552 ± 0.0172	**0.9748 ± 0.0014**	0.7884 ± 0.1108	**0.9735 ± 0.0013**
	Focal Loss	0.9744 ± 0.0023	0.9455 ± 0.0022	0.9411 ± 0.0035	0.9741 ± 0.0022	0.7224 ± 0.0298	0.9729 ± 0.0024
Mobile	Baseline	0.9745 ± 0.0021	0.9683 ± 0.0119	**0.9665 ± 0.0115**	0.9743 ± 0.0021	0.8263 ± 0.0502	0.9730 ± 0.0022
NetV3-L	Weighted Loss	0.9657 ± 0.0034	**0.9713 ± 0.0031**	0.9393 ± 0.0030	0.9705 ± 0.0026	**0.8853 ± 0.0524**	0.9638 ± 0.0036
ResNet50	Baseline	0.9682 ± 0.0050	0.9484 ± 0.0178	0.9494 ± 0.0196	0.9678 ± 0.0050	0.7890 ± 0.0611	0.9663 ± 0.0052
	Focal + Weighted Loss	0.9549 ± 0.0043	0.9520 ± 0.0181	0.9419 ± 0.0169	0.9552 ± 0.0044	0.7465 ± 0.0763	0.9524 ± 0.0045

The table reports classification results for EfficientNetV2-S, MobileNetV3-L, and ResNet50 under different training methods. Evaluation metrics include Accuracy, Balanced Accuracy, Macro F1, Weighted F1, Minority F1, and Cohen's κ. Values are reported as mean ± standard deviation across repeated experiments. The best result for each metric within a model is highlighted in bold.

However, because a two-stage classifier was not directly evaluated in this study, this interpretation should be regarded as a testable application arising from the observed trade-off rather than as a demonstrated improvement over a single-model system. Future studies could explicitly evaluate whether confidence-based routing between general and minority-focused classifiers improves specimen-level recognition in realistically imbalanced herbarium datasets.

4.3. Specimen-level recognition and implications for taxonomic reliability

The hierarchical evaluation framework adopted in this study provides a biologically meaningful assessment of model performance by evaluating recognition from cropped images to herbarium specimens and ultimately to species. Aggregating predictions from multiple cropped leaf images originating from the same specimen reduced the influence of isolated image-level errors, while species-level recognition was derived from the proportion of recognised specimens within each species. This framework more closely reflects specimen-based species

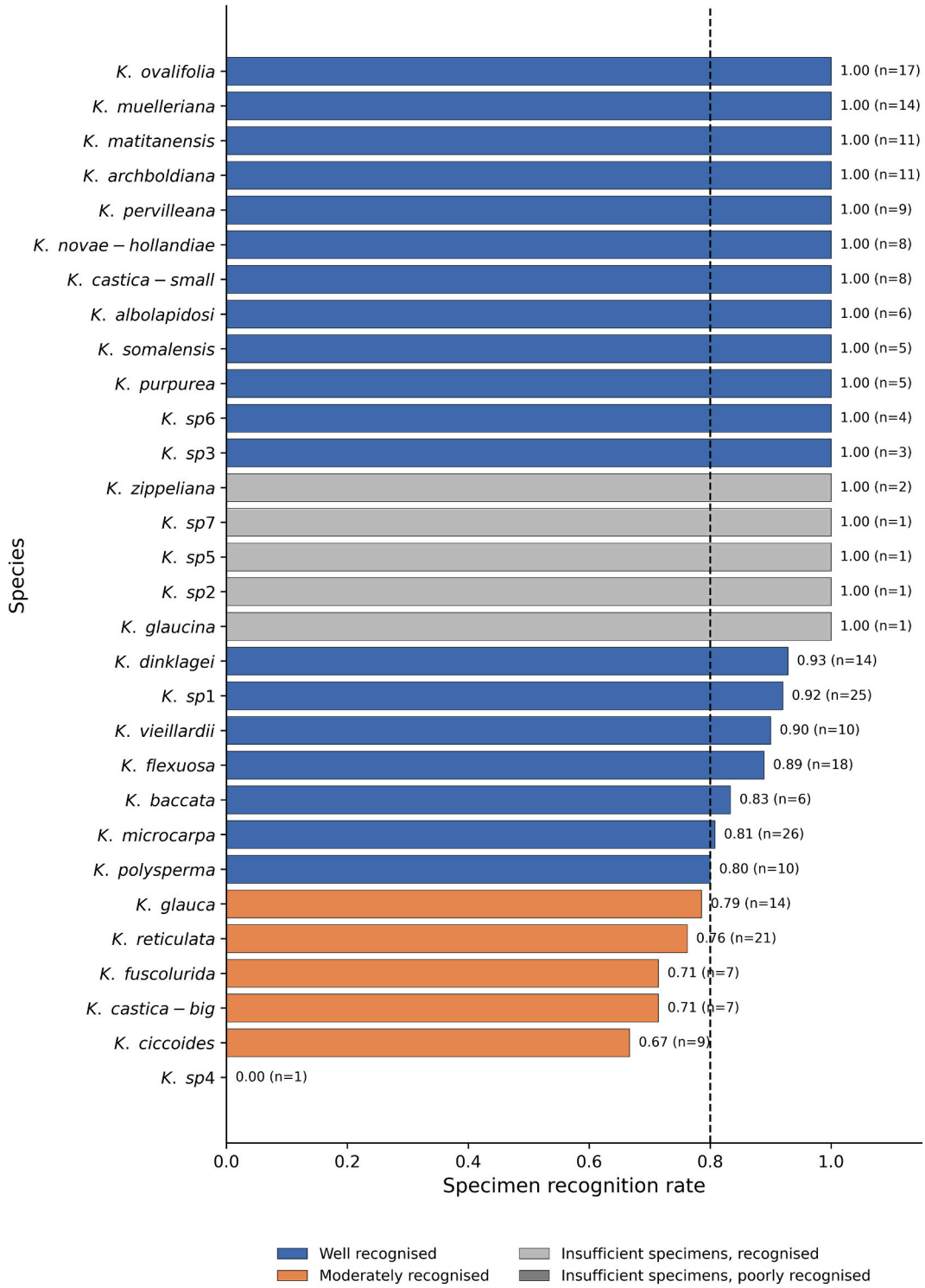


Fig. 11. Species-level specimen recognition rates across 30 *Kirganelia* species. Species represented by fewer than three specimens are shown descriptively as insufficiently sampled. The dashed line indicates the threshold for well-recognised species (recognition rate $\geq$ 0.80).

identification than image-level accuracy alone. When evaluated using this hierarchical framework, nearly 80% of sufficiently sampled *Kirganelia* species were classified as well recognised, whereas the remaining species were all moderately recognised. Notably, no sufficiently sampled species was classified as poorly recognised. In addition, reliable recognition was achieved for five of the six insufficiently sampled species. The only exception was *Kirganelia sp4*, for which the two misclassified cropped images consistently ranked the correct species as the second prediction and were confused only with the morphologically similar *K. albolapidosi* (despite differing in flowers, particularly longer stigmas). This restricted pattern of misclassification suggests that the model retained biologically meaningful taxonomic information despite extremely limited sampling.

The relatively small number of moderately recognised species provides further insight into the relationship between model performance and taxonomic complexity. Although these species did not reach the

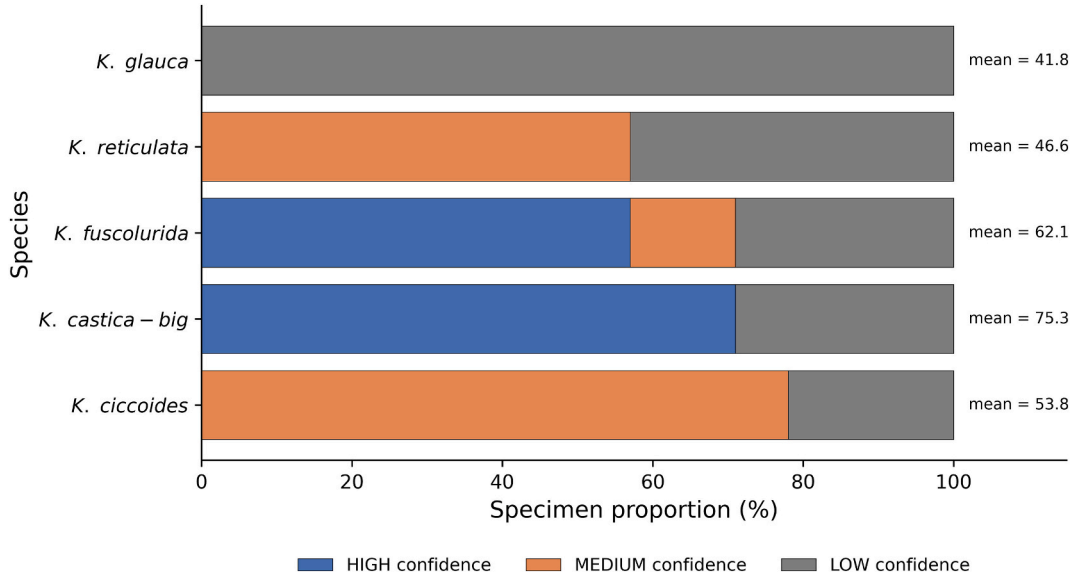


Fig. 12. Confidence-category composition among moderately recognised *Kirganelia* species. Stacked bars represent the proportion of HIGH, MEDIUM, and LOW confidence specimens for each species, with mean confidence values indicated beside bars.

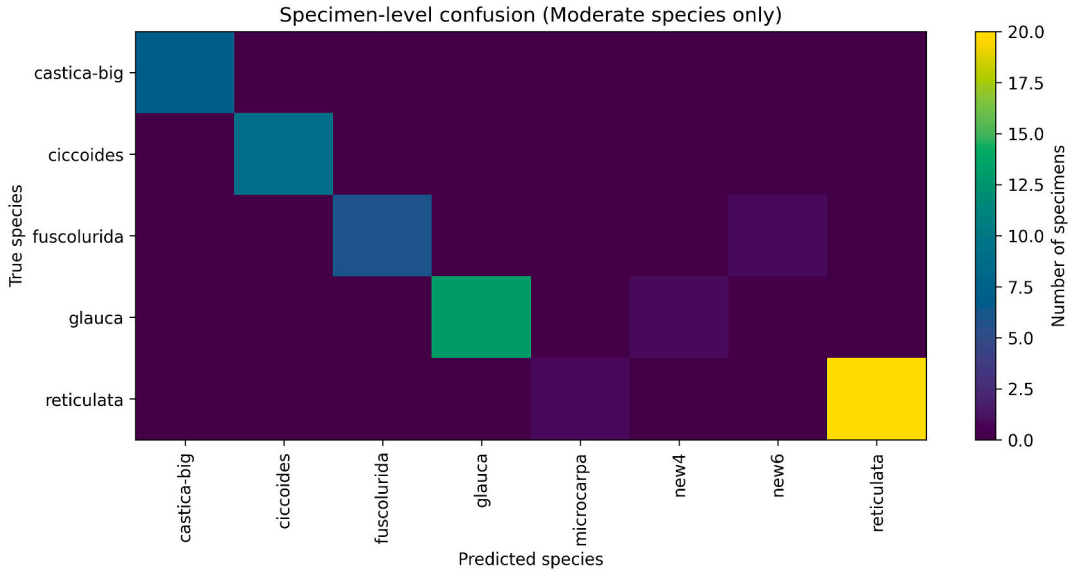


Fig. 13. Specimen-level confusion heatmap showing misclassification patterns among moderately recognised *Kirganelia* species. Rows represent true species labels and columns represent predicted species labels based on majority vote within folds and across folds. Color intensity indicates the number of specimens assigned to each predicted species.

threshold for well-recognised taxa, their specimen recognition rates fell within a narrow range (0.67–0.79), and all were still correctly identified for the majority of their specimens. Confidence analysis further revealed differences among species despite similar recognition rates. For example, *K. castica* was dominated by high-confidence specimens despite only moderate recognition, whereas *K. reticulata* and *K. glauca* showed lower prediction confidence. Notably, *K. castica* belongs to the Malagasy complex, which phylogenomic analyses recovered as non-monophyletic and affected by gene flow among lineages (Putri et al., 2026). High prediction confidence despite incomplete separability may therefore be consistent with genuine morphological overlap associated with a reticulate evolutionary history rather than solely reflecting a model error. Likewise, specimen-level confusion analysis showed that correct classifications remained dominant, with misidentifications restricted to relatively few specimens and only a small number of alternative species. Together, these findings indicate localized ambiguity among closely related taxa rather than systematic recognition failure.

These findings are consistent with previous studies showing that convolutional neural networks can successfully identify herbarium specimens despite the visual complexity of digitized collections, including variation in specimen preservation and the presence of labels and mounting materials (Carranza-Rojas et al., 2017; Unger et al., 2016; Younis et al., 2018). Developmental stage and other forms of intraspecific variation may further reduce the consistency of diagnostic characters, particularly among closely related species (Wäldchen and Mäder, 2018). However, unlike previous herbarium deep learning studies, particularly Shirai et al. (2022), which primarily emphasized image-level classification accuracy using large training datasets, the present study evaluated recognition at the specimen level within a taxonomically challenging lineage characterized by subtle interspecific

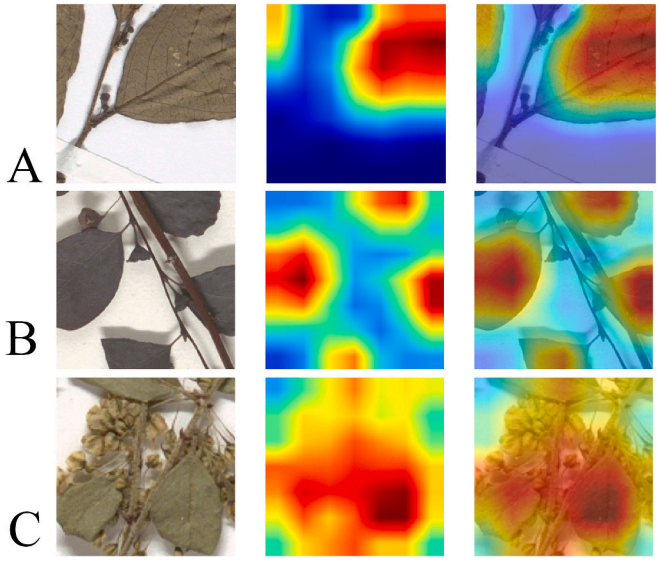


Fig. 14. Representative Grad-CAM visualizations illustrating the image regions contributing to species recognition by the best-performing CNN model. For each example, the original cropped image (left), Grad-CAM heatmap (middle), and heatmap overlay (right) are shown. (A) *Kirganelia ciccoides*; (B) *Kirganelia* sp. *6*. Both showing leaf-confined activation despite the presence of adjacent stems and reproductive structures. (C) *Kirganelia fuscolurida*, in which the activation extends to adjacent stem and reproductive structures while the dominant hotspot remains centered on leaf tissue.

differences, pronounced class imbalance, and limited specimen availability. Furthermore, species boundaries were established independently using phylogenomic evidence together with diagnostic reproductive morphology (Putri et al., 2026), allowing model performance to be interpreted within an evolutionary framework rather than solely through image similarity.

Another distinctive aspect of this study is the exclusive use of cropped leaf images. Because leaves are more consistently represented than reproductive structures in digitized herbarium collections, they provide the most widely available source of morphological information for large-scale image analysis. The high recognition achieved across most *Kirganelia* species demonstrates that vegetative morphology alone contains sufficient taxonomic information for reliable identification in many species. Conversely, the few moderately recognised taxa highlight cases in which reproductive characters may provide clearer taxonomic separation than leaf morphology. Rather than representing only technical classification errors, these cases may identify biologically informative instances where vegetative differentiation is limited despite independently supported species boundaries. By integrating hierarchical specimen-level evaluation, exclusive leaf-based image analysis, and phylogenomically validated species boundaries, this study demonstrates that deep learning performance can be interpreted within an evolutionary taxonomic framework rather than solely as an image classification task. Under this framework, patterns of correct recognition, confidence, and restricted misclassification provide biologically meaningful information about species distinctiveness and areas of morphological overlap. Consequently, the approach not only supports routine identification of herbarium specimens but also helps identify taxa that may warrant further morphological or phylogenomic investigation, highlighting the complementary roles of artificial intelligence and integrative taxonomy in biodiversity research (Karbstein et al., 2024).

4.4. Interpretation of CNN-based species recognition

Several limitations should be acknowledged when interpreting the present findings. First, the CNN models were trained exclusively on cropped leaf images. Although leaf morphology proved sufficient for accurate recognition of most *Kirganelia* species, incorporating additional plant organs or higher-resolution imaging of vegetative micromorphological features may further improve discrimination among morphologically similar taxa. Second, six species were represented by fewer than three specimens, preventing reliable species-level evaluation. Expanding sampling remains a priority, particularly where limited representation may reflect unresolved taxonomic diversity rather than biological rarity, as suggested by the phylogenomic separation of the two *K. castica* lineages (Putri et al., 2026).

Third, specimen and image characteristics may also contribute to variation in recognition performance. Herbarium images differed in preservation condition, mounting materials, and the amount of visible leaf tissue, while other properties such as image resolution, color balance, sharpness, and specimen age were not quantified systematically. Their individual effects therefore cannot be separated from biological variation in the present study. Nevertheless, the poor recognition of *Kirganelia sp4*, which was represented by limited material with overlapping leaves and extensive mounting tape, suggests that specimen presentation may affect classification in some cases. Future studies should explicitly evaluate whether specimen age, preservation quality, image resolution, color balance, image sharpness, and related imaging characteristics influence recognition performance.

Fourth, although Grad-CAM showed that predictions were consistently associated with leaf tissue, it could not identify the specific morphological characters underlying species recognition. Combining complementary explainable AI approaches with expert taxonomic assessment may therefore provide deeper biological interpretation of CNN decisions. Finally, because data partitioning was performed at the image-patch level rather than the specimen level, patches originating from the same herbarium specimen could be represented in different data splits. This may introduce some dependence between the training and evaluation data and slightly inflate performance estimates, although specimen-level aggregation and cross-validation were used to reduce this effect.

4.5. Limitations and future directions

Several limitations should be acknowledged when interpreting the present findings. First, the CNN models were trained exclusively on cropped leaf images. Although leaf morphology proved sufficient for accurate recognition of most *Kirganelia* species, incorporating additional plant organs or higher-resolution imaging of vegetative micromorphological features may further improve discrimination among morphologically similar taxa. Second, six species were represented by fewer than three specimens, preventing reliable species-level evaluation. Expanding sampling remains a priority, particularly where limited representation may reflect unresolved taxonomic diversity rather than biological rarity, as suggested by the phylogenomic separation of the two *K. castica* lineages (Putri et al., 2026). Third, although Grad-CAM showed that predictions were consistently associated with leaf tissue, it could not identify the specific morphological characters underlying species recognition. Combining complementary explainable AI approaches with expert taxonomic assessment may therefore provide deeper biological interpretation of CNN decisions. Finally, because data partitioning was performed at the image-patch level rather than the specimen level, patches originating from the same herbarium specimen could be represented in different data splits. This may introduce some dependence between the training and evaluation data and slightly inflate performance estimates, although specimen-level aggregation and cross-validation were used to reduce this effect.

4.6. Practical implementation in herbarium collections

Several barriers remain before CNN-based identification can be implemented routinely across herbarium collections. The first is the

availability of sufficiently curated training data. The present study used a taxonomically validated dataset in which species boundaries were independently supported and individual leaf images were extracted from herbarium specimens. Comparable datasets may not yet be available for many plant groups, particularly where specimen identifications remain uncertain or sampling is strongly imbalanced. For this reason, establishing reliable reference identifications and digitizing representative specimens may be more important initial requirements than increasing model complexity.

Computational resources represent a second practical constraint, particularly for smaller collections. Our results showed that performance gains beyond 20 epochs were limited, and that MobileNetV3-L remained competitive while providing strong minority-class recognition under weighted loss. This suggests that useful recognition performance does not necessarily require prolonged training or the most computationally demanding architecture. Transfer learning from pretrained CNNs, combined with targeted retraining on a local reference collection, may therefore provide a practical starting point for institutions with limited computing infrastructure.

The present results do not support a universal minimum number of specimens required per species. Although several species represented by only one or two specimens were recognised successfully, such cases cannot provide a reliable estimate of expected performance for other taxa. Recognition may also depend on class imbalance, morphological similarity, specimen quality, and image characteristics. Herbarium curators should therefore prioritize increasing taxonomic and morphological representation rather than aiming for a fixed numerical threshold. In practice, an initial system could be developed for better-represented taxa and progressively expanded as additional validated specimens become available, while low-confidence predictions and poorly represented species remain subject to expert verification. Under this framework, CNNs are best viewed as tools for prioritizing and supporting curatorial decisions rather than replacing taxonomic expertise.

5. Conclusions

This study demonstrates that CNN-based species recognition from leaf-based herbarium images can achieve high performance when combined with imbalance-aware training and a hierarchical evaluation framework that assesses recognition at the image, specimen, and species levels. Most evaluated *Kirganelia* species were accurately recognised despite substantial class imbalance, although the results also revealed a trade-off between maximizing overall classification performance and improving recognition of underrepresented species. Grad-CAM analysis further indicated that model predictions were primarily associated with leaf tissue rather than herbarium background. Together, these findings support the potential of deep learning as a biologically informed tool for herbarium species identification, while highlighting the importance of model selection, class-specific evaluation, and expert verification for poorly represented or uncertain taxa. Although these approaches cannot replace integrative taxonomy, they can provide a practical, non-destructive means of supporting specimen identification, prioritizing cases for expert review, and promoting broader use of digitized herbarium collections.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this manuscript, the authors used an AI-based language model to assist with improving clarity, organization, and academic writing style. All content was critically reviewed and edited by the authors, who take full responsibility for the accuracy and originality of the manuscript.

CRediT authorship contribution statement

Eva Kristinawati Putri: Writing – review & editing, Visualization, Validation, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Chiesa Serena:** Writing – review & editing, Visualization, Software, Methodology, Investigation, Formal analysis. **Lu Cao:** Writing – review & editing, Supervision, Methodology. **Hans ter Steege:** Writing – review & editing, Supervision. **Erick Paulus:** Writing – review & editing, Methodology. **Peter C. van Welzen:** Writing – review & editing, Supervision, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Eva Kristinawati Putri reports financial support was provided by Indonesia Endowment Fund for Education. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was conducted as part of the PhD project of Eva Kristinawati Putri and was supported by the Indonesia Endowment Fund for Education (LPDP, Scholarship No. 202203221508169). We gratefully acknowledge the following herbaria for supplying high-quality digitized images used in this study: A, BM, BR, E, G, K, L, MO, P, TI, U, US, and WAG (for abbreviations see https://sweetgum.nybg.org/science/ih/). We also thank the institutions and staff involved in the digitization of herbarium collections, whose efforts made this study possible. Computational analyses were performed using the Academic Leiden Interdisciplinary Cluster Environment (ALICE) provided by Leiden University.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecoinf.2026.104098.

Data availability

The code used for image preprocessing, model training, and evaluation is publicly available at: https://github.com/aseihcs/kirganelia-species-classification. The repository includes scripts for automatic leaf detection and cropping (get_cropped_images.py), as well as experiment pipelines for initial experiments, hyperparameter tuning, and final model training and evaluation.

References

Algemayel, O., Alhichri, H., Bazi, Y., Alajlan, N., Ammour, N., Alajlan, S., 2025. Deep learning for plant classification: a comprehensive review. Ecol. Inform. 77, 102368. https://doi.org/10.1016/j.ecoinf.2025.102368.

Ávila, M., Younis, S., Weiland, C., Seibold, S., Thines, M., Schmidt, M., Borsch, T., Müller, K.F., 2025. Herbariograph: a deep learning framework for species identification using herbarium specimens. New Phytol. 247, 1456–1470. https://doi.org/10.1111/nph.19742.

Baker, W.J., Bailey, P., Barber, V., Barker, A., Bellot, S., Bishop, D., Botigué, L.R., Brewer, G., Carruthers, T., Clarkson, J.J., Dodsworth, S., Epitawalage, N., Françoso, E., Gallego, B., Gottschling, M., Hipp, A.L., Hughes, C., Iganci, J., Johnson, M.G., Kim, J.T., Leempoel, K., Maurin, O., McGinnie, C., Pokorny, L., Roy, S., Scherberich, D., Schneider, J.V., Zuntini, A.R., Forest, F., 2022. A comprehensive phylogenomic platform for exploring the angiosperm tree of life. Syst. Biol. 71, 301–319. https://doi.org/10.1093/sysbio/syab035.

Bebber, D.P., Carine, M.A., Wood, J.R.I., Wortley, A.H., Harris, D.J., Prance, G.T., Davidse, G., Paige, J., Pennington, T.D., Robson, N.K.B., Scotland, R.W., 2010. Herbaria are a major frontier for species discovery. Proc. Natl. Acad. Sci. USA 107, 22169–22171. https://doi.org/10.1073/pnas.1011841108.

Belhumeur, P.N., Chen, D.M., Feiner, S., et al., 2008. Searching the world's herbaria: a system for visual identification of plant species. In: Computer Vision – ECCV 2008. Springer, Berlin, pp. 116–129. https://doi.org/10.1007/978-3-540-88693-8_9.

Bergstra, J., Bengio, Y., 2012. Random search for hyper-parameter optimization. J. Mach. Learn. Res. 13, 281–305. https://doi.org/10.5555/2188385.2188395.
Bonnet, P., Joly, A., Goëau, H., et al., 2018. Plant identification: man vs. machine. Multimed. Tools Appl. 77, 101–124. https://doi.org/10.1007/s11042-015-2607-4.
Bouman, R.W., Keßler, P.J.A., Telford, I.R.H., et al., 2022. A revised phylogenetic classification of tribe Phyllantheae (Phyllanthaceae). Phytotaxa 540, 1–76. https://doi.org/10.11646/phytotaxa.540.1.1.
Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M., 2010. The balanced accuracy and its posterior distribution. In: Proceedings of the International Conference on Pattern Recognition, pp. 3121–3124. https://doi.org/10.1109/ICPR.2010.764.
Buda, M., Maki, A., Mazurowski, M.A., 2018. A systematic study of the class imbalance problem in convolutional neural networks. Neural Netw. 106, 249–259. https://doi.org/10.1016/j.neunet.2018.07.011.
Canny, J., 1986. A Computational Approach to Edge Detection. https://doi.org/10.1109/TPAMI. 1986.4767851.
Carranza-Rojas, J., Goëau, H., Bonnet, P., et al., 2017. Going deeper in the automated identification of herbarium specimens. BMC Evol. Biol. 17, 181. https://doi.org/10.1186/s12862-017-1014-z.
Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20, 37–46. https://doi.org/10.1177/001316446002000104.
Cui, Y., Jia, M., Lin, T.-Y., et al., 2019. Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9268–9277. https://doi.org/10.48550/arXiv.1901.05555.
Dodsworth, S., Pokorny, L., Johnson, M.G., Kim, J.T., Maurin, O., Wickett, N.J., Forest, F., Baker, W.J., 2019. Hyb-Seq for flowering plant systematics. Trends Plant Sci. 24, 887–891. https://doi.org/10.1016/j.tplants.2019.07.007.
Eckert, I., Bruneau, A., Metsger, D.A., Joly, S., Dickinson, T.A., Pollock, L.J., 2024. Herbarium collections remain essential in the age of community science. Nat. Commun. 15, 7586. https://doi.org/10.1038/s41467-024-51899-1.
Guo, Y., Wang, X., Chen, C., Zhang, Y., Li, J., 2025. Applications of deep learning in herbarium specimen analysis: a review. Artif. Intell. Rev. 58, 102345. https://doi.org/10.1007/s10462-025-11408-2.
Haixiang, G., Li, Y., Shang, J., Mingyun, G., Yuanyue, H., Bing, G., 2017. Learning from class-imbalanced data: review of methods and applications. Expert Syst. Appl. 73, 220–239. https://doi.org/10.1016/j.eswa.2016.12.035.
Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., Fernández del Río, J., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E., 2020. Array programming with NumPy. Nature 585, 357–362. https://doi.org/10.1038/s41586-020-2649-2.
Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778. https://doi.org/10.48550/arXiv.1512.03385.
Hunter, J.D., 2007. Matplotlib: a 2D graphics environment. Comput. Sci. Eng. 9, 90–95. https://doi.org/10.1109/MCSE.2007.55.
Johnson, J.M., Khoshgoftaar, T.M., 2019. Survey on deep learning with class imbalance. J. Big Data 6, 27. https://doi.org/10.1186/s40537-019-0192-5.
Johnson, M.G., Gardner, E.M., Liu, Y., Medina, R., Goffinet, B., Shaw, A.J., Zerega, N.J. C., Wickett, N.J., 2019. HybPiper: extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. Appl. Plant Sci. 7, e11272. https://doi.org/10.1002/aps3.11272.
Kajihara, A.Y., Queiroz, G.A.D., Caxambu, M.G., et al., 2025. A database for automatic identification of herbarium specimens in Piperaceae. Multimed. Tools Appl. https://doi.org/10.1007/s11042-025-20883-2.
Karbstein, K., Kindt, R., Borsch, T., Boatwright, J.S., Cellinese, N., Fritsch, P.W., Grassa, C.J., Jost, L., Kadereit, G., Larridon, I., Muasya, A.M., Schönenberger, J., Simon, M.F., Smith, S.A., Soltis, D.E., Soltis, P.S., 2024. Artificial intelligence in integrative taxonomy and species delimitation. Trends Ecol. Evol. 39, 265–278. https://doi.org/10.1016/j.tree.2023.11.004.
King, G., Zeng, L., 2001. Logistic regression in rare events data. Polit. Anal. 9, 137–163. https://doi.org/10.1093/oxfordjournals.pan.a004868.
Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2980–2988. https://doi.org/10.1109/ICCV.2017.324.
Lorieul, T., Goëau, H., Bonnet, P., et al., 2019. Toward a large-scale and deep phenological stage annotation of herbarium specimens. Appl. Plant Sci. 7, e01233. https://doi.org/10.1002/aps3.1233.
Marín-Rodulfo, T., Soltis, P.S., Soltis, D.E., Smith, S.A., 2024. Old meets new: innovative and evolving uses of herbaria over time as revealed by a literature review. Plants People Planet 6, 123–137. https://doi.org/10.1002/ppp3.10530.
Marr, D., Hildreth, E.C., 1980. Theory of Edge Detection. https://doi.org/10.1098/rspb.1980.0020.
Mayo, S.J., Allkin, R., Baker, W.J., Blagoderov, V., Brake, I., Clark, B., Govaerts, R., Godfray, H.C.J., Haigh, A., Hand, R., Harman, K., Jackson, M., Kilian, N., Kirkup, D. W., Knapp, S., Malcolm, P., von Raab-Straube, E., Reynolds, C., Scott, B., White, R., Wilkinson, M., 2008. Alpha e-taxonomy: responding to the biodiversity crisis. Kew Bull. 63, 1–11. https://doi.org/10.1007/s12225-008-9013-2.
McKinney, W., 2010. Data structures for statistical computing in Python. In: Proceedings of the 9th Python in Science Conference, pp. 51–56. https://doi.org/10.25080/Majora-92bf1922-00a.
Meineke, E.K., Davies, T.J., Daru, B.H., Davis, C.C., 2020. Applying machine learning to investigate long-term insect–plant interactions preserved on digitized herbarium specimens. Appl. Plant Sci. 8, e11369. https://doi.org/10.1002/aps3.11369.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html.
Putri, E.K., Wielstra, B., Hendriks, K.P., van Welzen, P.C., Bouman, R.W., Brewer, R.F.A., de Oliveira, S.M., 2026. A phylogenomic framework for species delimitation in *Kirganelia* (Phyllanthaceae): tracing a history of reticulate evolution. Taxon 75, e70170. https://doi.org/10.1002/tax.70170.
Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 618–626. https://doi.org/10.1109/ICCV.2017.74.
Shirai, M., Takano, A., Kurosawa, T., Inoue, M., Tagane, S., Tanimoto, T., Koganeyama, T., Sato, H., Terasawa, T., Horie, T., Mandai, I., Akihiro, T., 2022. Development of a system for the automated identification of herbarium specimens with high accuracy. Sci. Rep. 12, 8066. https://doi.org/10.1038/s41598-022-11450-y.
Sokolova, M., Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. Inf. Process. Manag. 45, 427–437. https://doi.org/10.1016/j.ipm.2009.03.002.
Tan, M., Le, Q.V., 2021. EfficientNetV2: smaller models and faster training. In: Proceedings of the International Conference on Machine Learning, pp. 10096–10106. https://doi.org/10.48550/arXiv.2104.00298.
Unger, J., Merhof, D., Renner, S., 2016. Computer vision applied to herbarium specimens of German trees: testing the future utility of the millions of herbarium specimens for automated identification. BMC Evol. Biol. 16, 248. https://doi.org/10.1186/s12862-016-0827-5.
Wäldchen, J., Mäder, P., 2018. Plant species identification using computer vision techniques: a systematic literature review. Eco. Inform. 43, 1–12. https://doi.org/10.1007/s11831-016-9206-z.
Weitemier, K., Straub, S.C.K., Cronn, R.C., Fishbein, M., Schmickl, R., McDonnell, A., Liston, A., 2014. Hyb-Seq: combining target enrichment and genome skimming for plant phylogenomics. Appl. Plant Sci. 2, 1400042. https://doi.org/10.3732/apps.1400042.
Younis, S., Weiland, C., Schröder, S., et al., 2018. Taxon and trait recognition from digitized herbarium specimens using deep convolutional neural networks. Plant Methods 14, 102. https://doi.org/10.1080/23818107.2018.1446357.