www.nature.com/scientificdata

scientific data

OPEN

DATA DESCRIPTOR

BarkVisionAI: Novel dataset for rapid tree species identification

Ashwini Chhatre[1]✉, Nitesh Saini[1], Abhijeet Kumar Parmar[1], Pushpendra Rana[2] & Mayank Jain[3]

Tree species identification and mapping is crucial for forest management, biodiversity conservation, and ecological research. Bark images can be captured easily from the ground-level and can provide large amount of information about the tree species and its health. Yet, existing datasets for tree bark images are often limited in scope, lacking diversity in species representation and temporal attributes. To address these limitations, we present BarkVisionAI, a comprehensive dataset of 156001 tree bark images for 13 species collected from diverse forest types across India. Each image is labeled with location, species name, device attributes, and timestamp, providing a robust foundation for studying species identification and the variability of bark characteristics. We are providing detailed metadata information about each image, encouraging its use in ecological research, machine learning model training, and environmental monitoring. Benchmarking experiments using standard image classification models demonstrate the dataset's utility and effectiveness, highlighting its potential as a valuable resource for developing reliable, real-world applications in automated tree species identification and environmental change monitoring.

Background & Summary

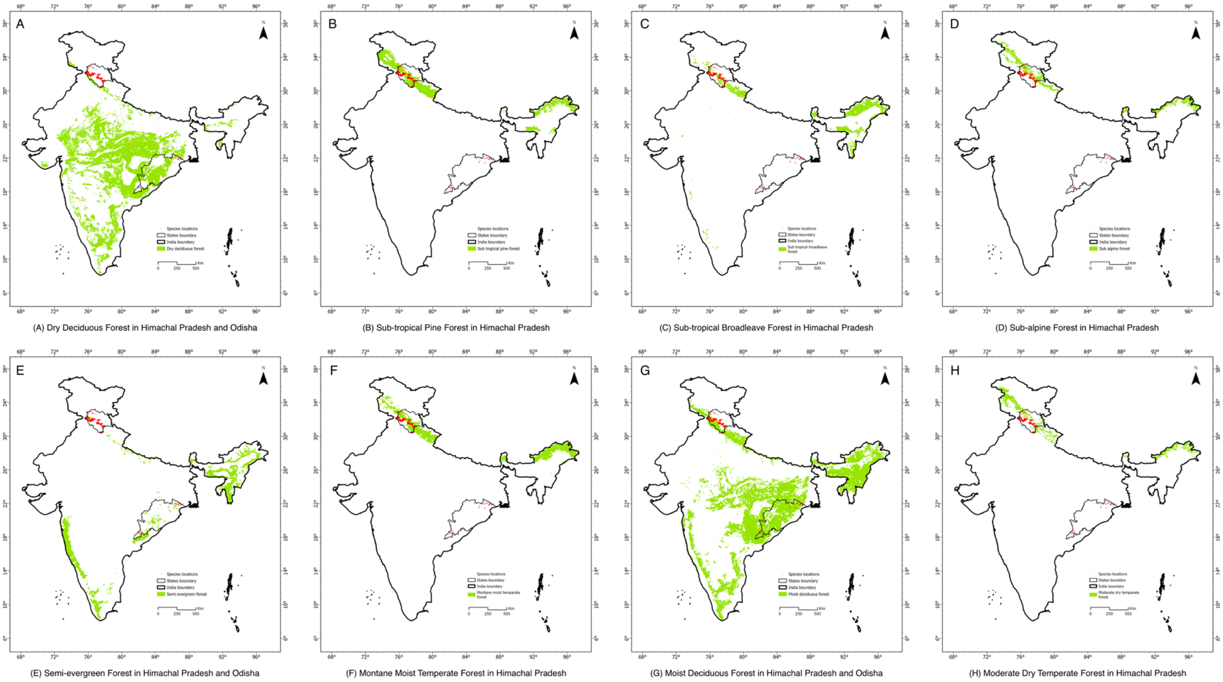
Tree species identification plays a pivotal role in numerous fields, including forestry, ecology, and conservation biology. Accurate identification is essential for forest management practices, biodiversity assessments, and monitoring ecological changes over time[1]. Among various morphological characteristics used for tree identification, bark texture and appearance provide significant, often underutilized, information[2,3]. Bark characteristics can be particularly useful in identifying species when leaves or other identifying features are not present, such as in winter or in heavily shaded forest environments[4].

Bark morphology can vary significantly between species and can also exhibit changes due to environmental factors such as climate, soil type, and weather conditions[5]. Additionally, the appearance of bark can change as climatic condition changes, making temporal variation important for accurate identification of tree species. While bark characteristics play a crucial role in tree species identification, many existing datasets primarily focus on leaves or other features, leading to a relative gap in comprehensive bark-related data. The BarkNet 1.0[6] dataset offers an important early benchmark for bark-based tree species classification, providing more than 23,000 high-quality images across 23 species and enabling systematic evaluation of model performance. Yet its usefulness beyond controlled settings is limited by its narrow geographic focus, modest species diversity, and relatively uniform imaging conditions, which together restrict how well models trained on it can adapt to the wide ecological and environmental variation found in real forests.

Consequently, it can be recognized that there is a clear need for a large, diverse dataset of tree bark images that covers a wide range of species and captures temporal and geographic variability. Such a dataset would be invaluable for developing automated systems for tree species identification, enhancing ecological research, and monitoring environmental changes. To address these gaps, we have created a comprehensive dataset of 156001 tree bark images from two dominant forest types across India. This dataset includes a diverse array of species from different ecological regions[7], providing a broad representation of India's rich forest biodiversity. Each image is labeled with the species name, camera device attributes, and timestamp, allowing for detailed analysis of species types and temporal changes in bark appearance.

Dataset overview. The dataset presented in this paper is the largest collections of tree bark images available, featuring 156001 images from a variety of forests across India. It includes over 13 different tree species,

[1]Bharti Institute of Public Policy, Indian School of Business, Hyderabad, India. [2]Himachal Pradesh Forest Department, Shimla, Himachal Pradesh, India. [3]Independent Researcher, Dublin, Ireland. ✉e-mail: ashwini_chhatre@isb.edu



(A) Dry Deciduous Forest in Himachal Pradesh and Odisha

(B) Sub-tropical Pine Forest in Himachal Pradesh

(C) Sub-tropical Broadleave Forest in Himachal Pradesh

(D) Sub-alpine Forest in Himachal Pradesh

(E) Semi-evergreen Forest in Himachal Pradesh and Odisha

(F) Montane Moist Temperate Forest in Himachal Pradesh

(G) Moist Deciduous Forest in Himachal Pradesh and Odisha

(H) Moderate Dry Temperate Forest in Himachal Pradesh

Fig. 1 The figure shows the distribution of the data across dominant forest types in India. The forest type layer is taken from Bhuvan Platform: https://bhuvan-app3.nrsc.gov.in/data/download/index.php?c=p&s=NICES&p=ftg&g=TS.

representing a wide range of ecoregions and forest types in India. The images are captured using standardized procedures to ensure high quality, and each image is tagged with precise species class label and timestamp data. This comprehensive metadata allows for detailed analyses of species types, environmental factors, and temporal changes in bark appearance.

To facilitate the use of this dataset, we have made available the entire dataset along with label and device metadata that will allow researchers, ecologists, and machine learning practitioners to access the images and metadata. The dataset is designed to support a wide range of applications, from developing machine learning models[6,8] for automated species identification to conducting ecological research on environmental monitoring[9].

Dataset significance and applications. The dataset serves as a valuable resource for several research domains. For ecologists and conservationists, it provides a means to identify tree species rapidly and study the temporal changes in tree bark which is effective in monitoring environmental changes over time. For the machine learning community, the dataset offers a benchmark for developing and testing image classification models in real-world conditions, with diverse species and temporal data that reflect the complexity of natural environments.

Our initial experiments using standard image classification models have demonstrated the effectiveness of the dataset in training models to accurately identify tree species based on bark images. These results highlight the potential of the dataset to advance research in automated tree species identification, environmental monitoring, and biodiversity conservation. This paper aims to provide a detailed description of the dataset, including the methods used for data collection and annotation, the structure and format of the data, and the results of our benchmark experiments.

Methods

Dataset collection. The site selection and data collection process for the BarkVisionAI dataset was a comprehensive exercise involving detailed planning, stakeholder coordination, and systematic implementation. The BarkVisionAI data represents dominant tree diversity in India across 8 dominant forest types (Fig. 1) and 9 ecoregions[7] (Figs. 2,3). The state of Himachal Pradesh and Odisha were selected to ensure the maximum forest type coverage and diversity of ecoregion.

Himachal pradesh data collection exercise. The data collection exercise was launched through a planning meeting that brought the Himachal Pradesh Forest Department (HPFD) as key stakeholders. The focus of this meeting was to finalize the roadmap, identify representative species, and select forest divisions. The data collection exercise targeted the lesser and mid-Himalayan zones, spanning across six forest divisions (Anni, Dehra, Nachan, Nurpur, Palampur, and Paonta Sahib). Species of ecological and economic importance were prioritized based on their abundance in the selected areas. Specific tasks, such as data collection by forest guards, were planned in detail, with HPFD anchoring the operational aspects and authors providing technical and training support. The process also included the generation of training materials, defining minimum data requirements,

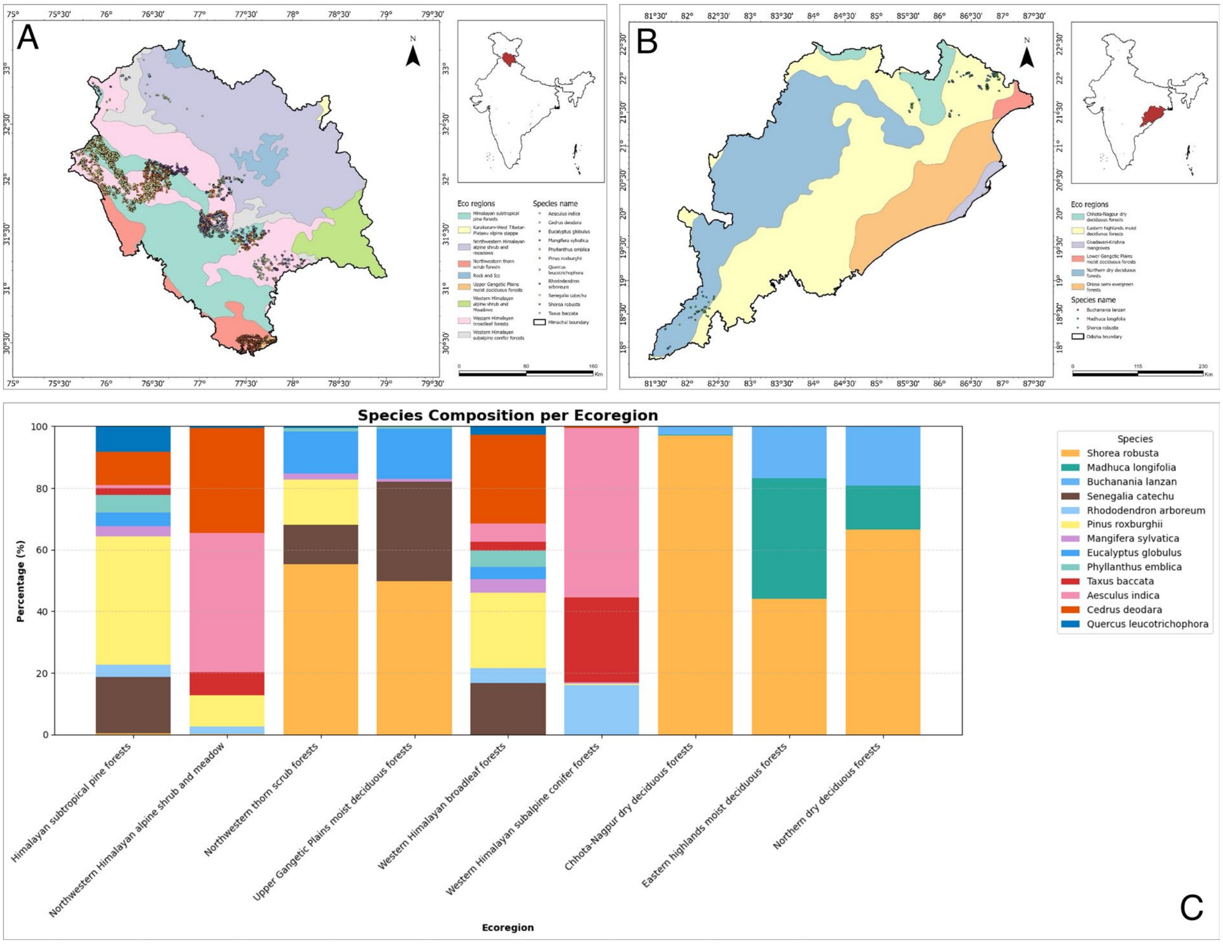


Fig. 2 (**A,B**) Maps showing distribution of tree species across eco-regions in Himachal Pradesh and Odisha in India. The ecoregion boundary data is taken from Data Basin: https://databasin.org/datasets/68635d7c77f1475f9b6c1d1dbe0a4c4c/. (**C**) The stacked percentage chart shows the relative composition of species within each ecoregion.

and establishing a phased approach to data validation and monitoring. The involvement of divisional officers ensured alignment with local priorities and facilitated the inclusion of species deemed important by field officers.

Data collection began with an intensive training program for forest staff, including forest guards, range officers, and divisional forest officers. The training sessions introduced participants to the digital data collection platform and its mobile applications (Open Data Kit) for recording data. Practical demonstrations covered aspects such as location accuracy (5–10 m), and image capture requirements (distance from tree, specific angles and zoom levels).

The process was supported by dynamic monitoring mechanisms, including a real-time dashboard that tracked progress across divisions and species. Challenges such as poor network connectivity in remote areas were mitigated through preemptive planning, like using local Wi-Fi for registrations and alternate training venues. Stakeholder feedback was incorporated iteratively, leading to updates in training materials and data capture forms. Continuous hand-holding support through messaging groups, field visits, and regular reviews ensured the quality and accuracy of the collected data.

Odisha data collection exercise. The data collection exercise in Odisha was conducted by a team of experts appointed by the authors, ensuring adherence to a systematic and standardized methodology. Drawing from the protocol established in Himachal Pradesh, the Odisha data collection drive incorporated specific adaptations to address the region's distinct ecological and geographical characteristics. The region is dominated by tropical moist and dry deciduous forest. Key steps included the selection of sampling sites, ensuring representation across diverse forest types and habitat conditions prevalent in the state. Detailed procedures were implemented for species identification and data recording. Comprehensive training sessions were organized for field staff, emphasizing the accuracy of image capture.

Image acquisition. Tree bark image acquisition using mobile phone cameras was conducted under a standardized protocol to ensure consistency and precision. High-resolution smartphones with cameras of at least 8 MP were employed, capturing images at varying distances ranging from 50 cm to 2 meters to emphasize bark texture

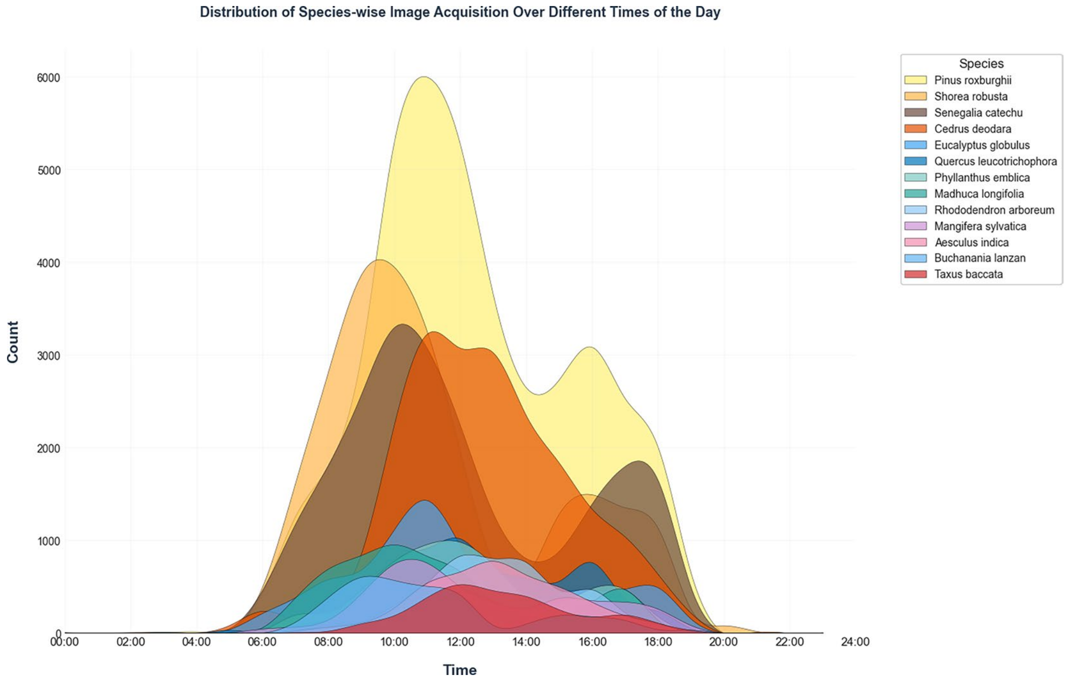


Fig. 3 The distribution of image acquisition during different time of the day.

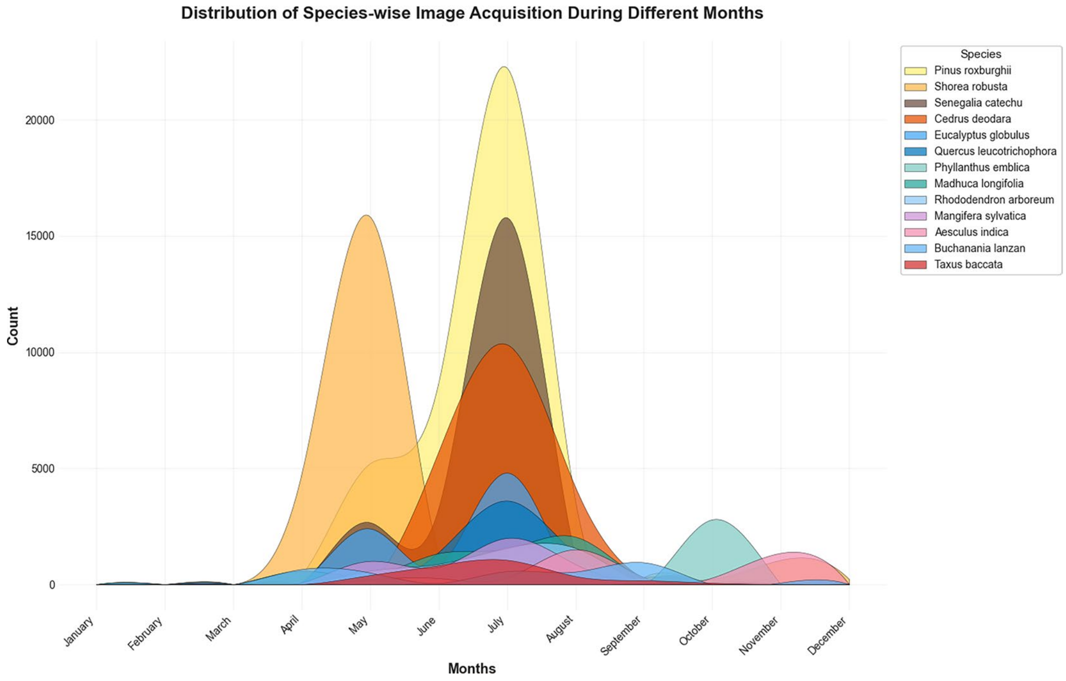


Fig. 4 The distribution of image acquisition during different months.

and patterns. The dataset includes images captured using 315 unique camera models from 20 different mobile phone manufacturers, ensuring diverse device representation. In most instances, phones were held perpendicular to the bark surface to minimize distortion, and images were taken under varying lighting conditions at different times of the day to enhance dataset diversity (Fig. 3). Data collection spanned January 2024 to December 2024, capturing temporal variations across summer, monsoon, and winter seasons (Fig. 4). This temporal variability provides insights into seasonal changes in bark conditions, facilitating the development of machine learning models for ecological analyses. Special emphasis was placed on recording distinctive features such as cracks, ridges, and peeling layers to aid in accurate species identification. GPS coordinates and timestamps were automatically embedded with each image, ensuring precise spatial and temporal metadata. A dedicated portal was created to perform the random tree species label validation by forestry experts (https://validate.ncount.in). This

Species name	Number of images
Pinus roxburghii	40451
Shorea robusta	24968
Senegalia catechu	23030
Cedrus deodara	20633
Eucalyptus globulus	8204
Quercus leucotrichophora	6485
Phyllanthus emblica	6262
Madhuca longifolia	6084
Rhododendron arboreum	4719
Mangifera sylvatica	4714
Aesculus indica	4233
Buchanania lanzan	3363
Taxus baccata	2855
Total	156001

Table 1. Number of image records for each species type in the dataset.

systematic approach resulted in a robust dataset of high-quality bark images, suitable for species identification and broader ecological research applications.

Data Records

Table 1 provides a detailed breakdown of the number of images captured for each species in the dataset, highlighting the representation of diverse taxa.

The data records consist of comprehensive information captured during field data collection, combining both visual and contextual metadata for each observation. The images for each species type are stored in separate folders. The folders are named as per the scientific name of the species. Each record includes high-resolution images, timestamps, and species identification details. Associated metadata encompasses image and camera attributes. The metadata records (*BarkVisionAI_metadata.csv*) are provided as separate file inside the parent dataset folder. The records are structured and standardized formats are employed for consistency as specified in Table 2.

Data directory structure. The BarkVisionAI dataset has been uploaded to Figshare in zipped file format[20] Fig. 5 depicts the directory structure of the BarkVisionAI dataset. The zipped dataset file contains a root directory that branches into two main subdirectories: *Balanced_Data*, and *BarkVisionAI*. Each of these two subdirectories follows an identical internal structure, containing its own respective metadata file in CSV format (e.g., *Balanced_Data_Metadata.csv*, *BarkVisionAI_Metadata.csv*) and a *data/* folder. Within this *data/* folder, the image files are further organized into 13 tree species-specific subfolders, including *Aesculus indica/*, *Buchanania lanzan/, Cedrus deodara/, Shorea robusta/*, and *Taxus baccata/*, among others. The individual image files in JPEG format such as 1714034904493.jpg, are stored within their corresponding species folder.

Technical Validation

Creation of balanced dataset. To test the applicability of BarkVisionAI data for automated species classification, a balanced and unbiased dataset is required. A dataset generated through simple random sampling would likely contain significant confounding variables[10], where spurious correlations—such as a specific species being photographed predominantly with a single device or during a single season—could be learned by a model instead of the intrinsic, morphological features of the bark.

To mitigate these risks, a stratified sampling strategy was implemented. The objective of this strategy was to produce a final dataset that is balanced across three key contexts: environmental conditions, temporal contexts, and technical (device) characteristics. These contexts were captured by five critical covariates: elevation gradients, seasonal phenological changes, diurnal lighting conditions, seasonal timing, and image capture device variability.

This stratification serves as a multi-dimensional methodological control. The temporal and phenological variables act as proxies for photometric conditions (i.e., illumination variance), while the device variable controls for sensor-specific characteristics (i.e., sensor bias). By ensuring each species is represented across these varying conditions, the resulting dataset is designed to de-correlate the species label from data collection artifacts, thereby facilitating the development of models that learn true biological features.

The stratification process was applied to a pre-filtered data pool. The total metadata database contained 167,361 records. However, 11,360 of these records were missing the location information necessary for deriving elevation data. These records were excluded, establishing a usable data pool of 156,001 records from which the stratified selection was performed. For users who wish to work exclusively with species-labelled images without any accompanying contextual metadata, we have made the complete set of raw images available on Zenodo, accessible at: https://doi.org/10.5281/zenodo.14650999.

Derivation of environmental and phenological covariates. Prior to sampling, four new covariates were derived from the existing metadata to serve as the basis for stratification: *elevation_level, leaf_status, time_of_day*, and *month*.

Metadata field	Description
image_id	Unique identifier for the image
species_name	Scientific name of the species
latitude	Latitude of the tree
longitude	Longitude of the tree
altitude	Altitude/elevation of the location where coordinates are captured
altitude_dem	Elevation value of the coordinate location derived from the SRTM 30 m DEM
accuracy	Positional accuracy as reported by ODK application
state_name	Name of the state in India where data is captured
phone_make	Manufacturer of the mobile phone
phone_camera_model	Camera model specified by the Manufacturer in the technical specification document of the mobile phone
encoding_process	Method used to compress and encode image data
rear_main_camera_megapixels	Megapixel value specified by the Manufacturer in the technical specification document of the mobile phone
rear_main_camera_aperture_size	Aperture size specified by the Manufacturer in the technical specification document of the mobile phone
rear_camera_flash_available	Availability of flash in the rear camera. Binary value indicated as Yes or No
original_image_size	Original image size captured by the mobile phone camera as recorded in the image metadata
original_image_resolution	Original image megapixel resolution captured by the mobile phone camera as recorded in the image metadata
processed_image_size	Image size after cropping operation
image_date_time	Image acquisition date and time as recorded in the image metadata
elevation_level	Species-specific elevation level. Indicated altitude range of natural habitat of the species.
leaf_status	Field indicating leaf-on or leaf-off status for trees. Evergreen tree species have leaf-on status.
time_of_day	Time of image acquisition as per following classification: [Morning: 5:00 Hrs to 11:00 Hrs; Afternoon: 11:01 Hrs to 15:00 Hrs; Evening: 15:01 Hrs to 19:00 Hrs.]
month	Month of the image acquisition

Table 2. Description of image metadata fields included in the dataset records.

Altitudinal distribution (elevation_level). To capture environmental diversity[11] related to altitudinal gradients, an *elevation_level* column was derived from the *altitude_dem* metadata field. The derivation was not based on global, fixed-altitude bins, but was instead a relative, per-species calculation. The process involved grouping the dataset by species name. For each species group containing at least five records, quantile-based binning (via pd.qcut function in Pandas library of Python) was applied to its *altitude_dem* values. This method discretizes the elevation data into five equal-frequency bins, or quantiles, labeled *level1, level2, level3, level4*, and *level5*.

The significance of this per-species, quantile-based approach is that it transforms an absolute physical measurement (meters above sea level) into a relative ecological proxy. For example, level1 for a high-altitude species like *Cedrus deodara* (observed range 500–3200 m) and level1 for a low-altitude species *Buchanania lanzan* (observed range 40–1200 m) both represent the lower 20th percentile of that specific species' observed altitudinal range. This stratification ensures that the final dataset for each species contains a balanced number of samples from the low-, mid-, and high-elevation extents of its habitat, compelling a model to learn features that are invariant to altitude-related environmental pressures.

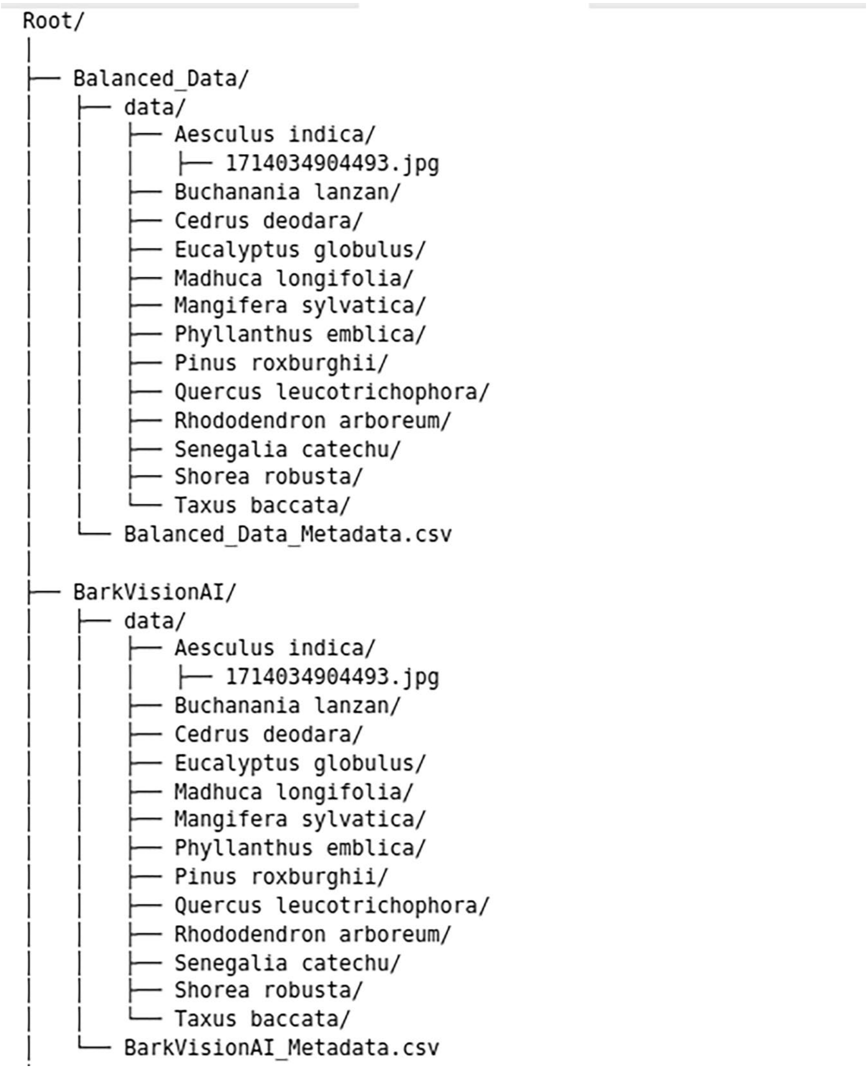
Seasonal phenology (leaf_status). To account for significant seasonal changes in tree morphology and, critically, in the lighting conditions on the trunk, a leaf_status column was derived. This variable indicates whether a tree was in a *leaf_on* or *leaf_off* condition at the time of image capture. The derivation was managed by a species-specific, rule-based phenological dictionary that maps month numbers to leaf status based on known biological behavior. For each record, a *month_num* (integer 1–12) was extracted from its image timestamp. This month was then compared against the dictionary rules for that record's species.

The dictionary and logic handled two distinct species types:

- Evergreen species (e.g., *Cedrus deodara, Pinus roxburghii, Taxus baccata*) were invariably labeled *leaf_on*.
- Deciduous species (e.g., *Aesculus indica, Shorea robusta*) had explicitly defined *leaf_on* and *leaf_off* months.

The complete phenological dictionary used for this derivation is provided in Table 3. This variable serves as a direct proxy for canopy cover, which dictates whether the bark is illuminated by diffuse, shaded light (*leaf-on*) or direct, hard light (*leaf-off*).

Derivation of temporal and technical covariates. Temporal and lighting conditions (time_of_day, month). In addition to the phenological *leaf_status*, two other temporal variables were derived from the *image_date_time* metadata column: *month and time_of_day*. The month column was extracted as an integer (1–12) from the image timestamp. Its primary purpose was to enable seasonal interpretation and to serve as the input for the *leaf_status* derivation. The



```
Root/
|
├── Balanced_Data/
|    ├── data/
|    |    ├── Aesculus indica/
|    |    |    ├── 1714034904493.jpg
|    |    ├── Buchanania lanzan/
|    |    ├── Cedrus deodara/
|    |    ├── Eucalyptus globulus/
|    |    ├── Madhuca longifolia/
|    |    ├── Mangifera sylvatica/
|    |    ├── Phyllanthus emblica/
|    |    ├── Pinus roxburghii/
|    |    ├── Quercus leucotrichophora/
|    |    ├── Rhododendron arboreum/
|    |    ├── Senegalia catechu/
|    |    ├── Shorea robusta/
|    |    └── Taxus baccata/
|    └── Balanced_Data_Metadata.csv
|
├── BarkVisionAI/
|    ├── data/
|    |    ├── Aesculus indica/
|    |    |    ├── 1714034904493.jpg
|    |    ├── Buchanania lanzan/
|    |    ├── Cedrus deodara/
|    |    ├── Eucalyptus globulus/
|    |    ├── Madhuca longifolia/
|    |    ├── Mangifera sylvatica/
|    |    ├── Phyllanthus emblica/
|    |    ├── Pinus roxburghii/
|    |    ├── Quercus leucotrichophora/
|    |    ├── Rhododendron arboreum/
|    |    ├── Senegalia catechu/
|    |    ├── Shorea robusta/
|    |    └── Taxus baccata/
|    └── BarkVisionAI_Metadata.csv
|
```

Fig. 5 The data directory structure of BarkVisionAI dataset.

time_of_day column was derived from the hour of capture and categorized into three specific bins to represent distinct lighting conditions:

- Morning (5:00 Hrs – 11:00 Hrs)
- Afternoon (11:01 Hrs – 15:00 Hrs)
- Evening (15:01 Hrs – 19:00 Hrs)

These temporal bins are not uniform; they are photographically significant. The 'Afternoon' bin corresponds to the period of highest solar altitude, which typically produces high-contrast, "flat" illumination that can obscure 3D texture. Conversely, the 'Morning' and 'Evening' bins capture lower-angle, "raking" light, which is known to accentuate the 3D surface detail and texture of the bark. Stratifying on these bins ensures the dataset contains a balanced representation of images with both flat and textured illumination profiles. By stratifying on both *month* and *leaf_status*, the sampling algorithm can de-correlate seasonal effects (e.g., low sun angle in winter) from phenological effects (e.g., no canopy cover). For example, the system can balance the inclusion of *month = 1* records for *Shorea robusta* (which is leaf_on) against *month = 1* records for *Phyllanthus emblica* (which is *leaf_off*), preventing the model from learning a simple, incorrect heuristic for "winter".

Device and sensor variability (phone_camera_model). The fifth stratification variable, *phone_camera_model*, was already available in the collected metadata. This column identifies the specific mobile device used to capture each image.

Inclusion of this variable is a direct countermeasure against a common failure mode in deep learning known as "shortcut learning[12]." Without this control, a model could learn to identify a species by spuriously correlating it

Species Name	Leaf-on Months	Leaf-off Months
Aesculus indica	2, 3, 4, 5, 6, 7, 8, 9, 10	11, 12, 1
Buchanania lanzan	2, 3, 4, 5, 6, 7, 8, 9, 10	11, 12, 1
Cedrus deodara	Evergreen	—
Eucalyptus globulus	Evergreen	—
Madhuca longifolia	2, 3, 4, 5, 6, 7, 8, 9, 10, 11	12, 1
Mangifera sylvatica	3, 4, 5, 6, 7, 8, 9, 10, 11	12, 1, 2
Phyllanthus emblica	3, 4, 5, 6, 7, 8, 9, 10, 11, 12	1, 2
Pinus roxburghii	Evergreen	—
Quercus leucotrichophora	Evergreen	—
Rhododendron arboreum	1, 2, 3, 4, 5, 6, 7, 8, 9, 12	10, 11
Senegalia catechu	1, 5, 6, 7, 8, 9, 10, 11, 12	2, 3, 4
Shorea robusta	1, 5, 6, 7, 8, 9, 10, 11, 12	2, 3, 4
Taxus baccata	Evergreen	—

Table 3. Species-specific phenological dictionary used for the derivation of the *leaf_status* covariate, mapping calendar months to the "*leaf-on*" or "*leaf-off*" state based on known phenological patterns.

with the unique sensor noise, color science, or post-processing artifacts of a specific phone brand that was disproportionately used to photograph that species. By stratifying on *phone_camera_model*, the dataset ensures that the sample set for each species contains, as much as possible, a representative mix of images from all available device types. This renders the camera model a useless predictor, forcing the classifier to learn the true, generalizable morphological features of the bark that are invariant across different capture devices and image quality profiles.

Stratified sampling algorithm and execution. Stratification schema construction. To execute the multi-dimensional stratification, a new composite column, *stratify_col*, was constructed for each record. This column was created by concatenating the values of the five critical attributes:

1. *phone_camera_model*
2. *time_of_day*
3. *month*
4. *leaf_status*
5. *elevation_level*

Each unique value in this *stratify_col* represents a single, fine-grained stratum (e.g., a combination like "iPhone12-Morning-3-leaf_on-level2").

Per-Species proportional sampling and contingency. The sampling algorithm was applied on a per-species basis to ensure a balanced class distribution in the final dataset. The procedure was as follows:

1. The 156,001-record dataset was grouped by species name.
2. Within each species group, the frequency of each unique stratify_col (stratum) was identified.
3. The primary sampling method was to sample proportionally from each stratum, without replacement. This ensures that the representation of each stratum in the final dataset is relative to its frequency in the original data pool.
4. The sampling was capped to a target size of n = 2800 samples per species.
5. A contingency was defined to handle the sparse nature of high-dimensional ecological field data: If the proportional sampling process failed to reach the target of 2,800 samples (due to many underrepresented or empty strata), the remaining slots were filled by taking additional random samples from the remaining, non-selected records within that species group.

This two-step hybrid algorithm was a pragmatic solution that prioritized the most critical objective for classifier training: achieving a perfectly balanced dataset with 2,800 samples per class. The "*proportional*" first pass maximized the environmental, temporal, and technical diversity of this set, while the "*random-fill*" contingency ensured the target sample size was met.

Final dataset characteristics. The resulting dataset, named *stratified_selection*, contains a total of 36,400 samples. This total reflects a perfectly balanced class distribution of 2,800 samples for each of the 13 species represented in the dataset (36,400 total samples/2,800 samples per species = 13 species). This final dataset achieves the study's objective by providing a balanced set of samples for each species. This balance is not only numerical (i.e., 2,800 samples per class) but also contextual, as each species' dataset is internally stratified to capture maximum

Model	Training Accuracy on BarkVisionAI dataset	Validation Accuracy on BarkVisionAI dataset	Hyperparameter and model attributes
ResNet18	86.47	84.90	• Number of Epochs: 20 • Batch Size: 64 • Learning Rate: 0.0001 • Weight Decay: 0.0001 • Training Split Ratio: 0.6375 • Validation Split Ratio: 0.1125 • Testing Split Ratio: 0.25 • Learning Rate Decay: 0.2 • Optimizer: Adam • Loss Function: CrossEntropyLoss • Evaluation Metrics: Confusion Matrix • Number of Classes: 13 • Total Images per Class: 2800 • Image Size: 512 × 512
ResNet34	87.45	86.42	
ResNet50	88.18	87.42	
VGG16	80.85	80.45	
EfficientNetB0	84.44	83.96	
NvidiaEfficientNetB4	73.38	72.17	
ViT_base_patch14_ reg4_dinov2.lvd142m (PlantClef 2024 model)	84.84	85.03	• Number of Epochs: 50 • Batch Size: 16 • Learning Rate: 0.0001 • Weight Decay: 0.0001 • Training Split Ratio: 0.6375 • Validation Split Ratio: 0.1125 • Testing Split Ratio: 0.25 • Learning Rate Decay: 0.2 • Optimizer: Adam • Loss Function: CrossEntropyLoss • Evaluation Metrics: Confusion Matrix • Number of Classes: 13 • Total Images per Class: 2800 • Image Size: 512 × 512

Table 4. Training and validation accuracies of different deep learning models on the balanced BarkVisionAI dataset (13 classes, 2800 images); with key training hyperparameters.

diversity across the selected key factors. This construction method ensures the dataset's suitability for training and validating generalizable, artifact-free species classification models.

Experimental validation and results. To demonstrate the technical efficacy of the BarkVisionAI dataset, a series of AI-based classification tests were conducted. The dataset, comprising 13 classes with 2,800 images each, was split into training (63.75%), validation (11.25%), and testing (25.0%) sets. All images were processed at 512 × 512 resolution. A range of established convolutional neural network (CNN) architectures were benchmarked, along with a state-of-the-art Vision Transformer (ViT) model, which was a baseline model for the PlantCLEF 2024 challenge[13]. The validation accuracies for the tested models are summarized in Table 4. The ResNet50 architecture achieved the highest validation accuracy, slightly outperforming the more complex ViT model.

Given its top-performing status, the ResNet50 model[14] was analyzed in further detail. On the 9,100-sample test set, the model achieved an overall accuracy of 87.42%, correctly identifying 7,956 images. A breakdown of its performance against the dataset's stratification variables revealed the impact of the controlled covariates (Fig. 6).

The model's misclassification rate showed a distinct correlation with the time of the day. The highest rate was observed during the "Evening," while the "Morning" and "Afternoon" periods had consistently lower and similar misclassification rates. This suggests that the diminished natural light during evening hours, even when balanced, impacts image quality and feature clarity, leading to more prediction errors. The model performed most reliably in stable lighting conditions. Performance also varied by the altitude gradient. The model performed best at low altitudes, with misclassification rates increasing at medium and high elevations. This may indicate that greater environmental variability or potential for species overlap at higher elevations introduces additional visual complexity, making classification more challenging.

A combined analysis showed that accuracy varied across both months and times of day:

1. Peak performance was noted in September, particularly during the morning and evening (both 92% accuracy).
2. Lighting conditions appeared critical, as the "Afternoon" (11:00 Hrs - 15:00 Hrs) slot generally resulted in higher accuracy across most months, likely due to consistent, high-angle illumination.
3. Lower performance was seen in specific strata, such as October mornings (79%) and April evenings (81%). These drops may be linked to a relatively smaller number of images within those specific, fine-grained strata.
4. Data Anomaly: A 100% accuracy was recorded for December, but this is statistically insignificant as it was based on an extremely small sample size ($n=6$) in the test set.

The experimental results successfully validate the technical efficacy of the stratified dataset. The high overall accuracy of 87.42% achieved by the ResNet50 model demonstrates that the balanced dataset enables a classifier to learn generalizable features of tree bark for species identification. The detailed performance analysis further confirms the necessity of the stratification strategy. The observed variations in model accuracy across different times of day (lighting), elevations (environmental proxies), and months (seasonal/phenological changes)

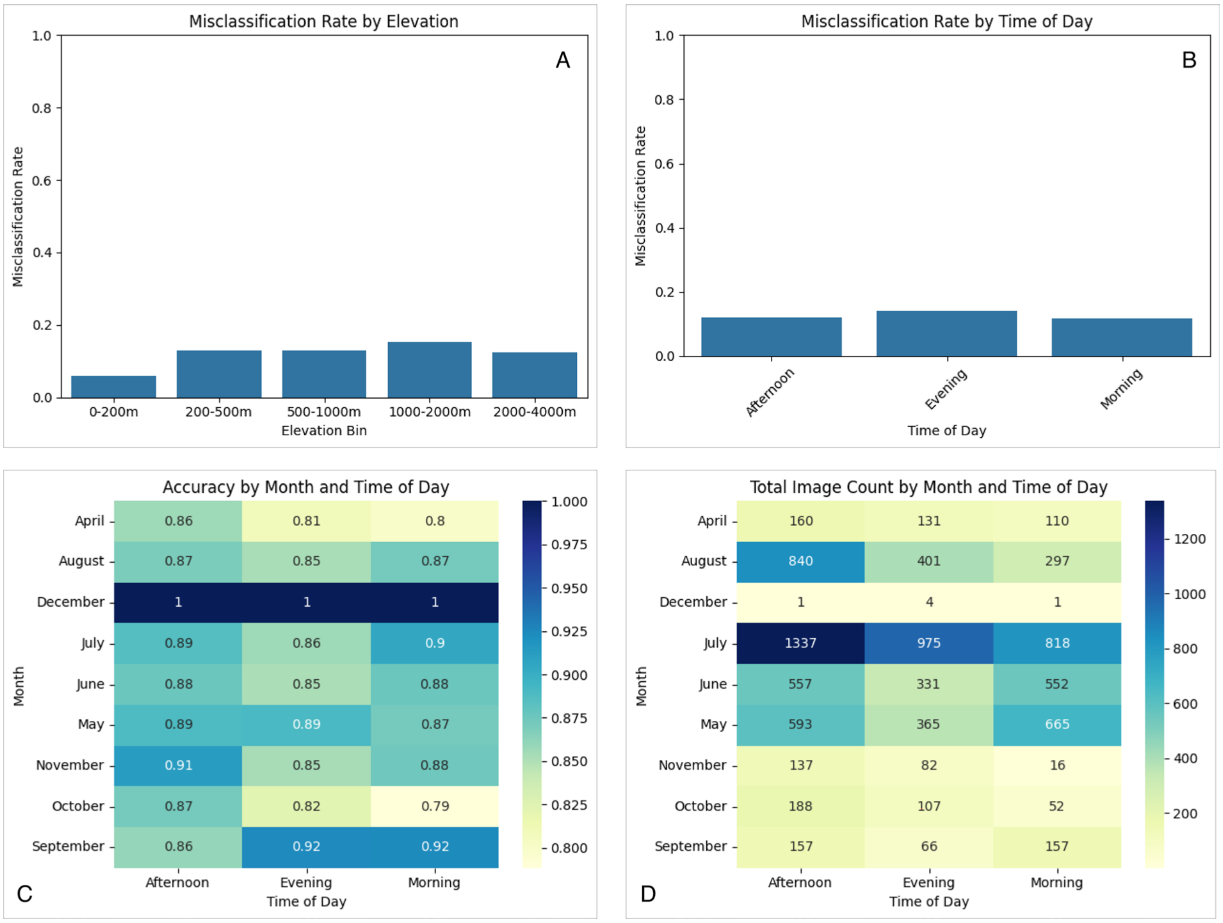


Fig. 6 (**A,B**) Resnet50 misclassification rate by elevation and time of the day. (**C**) Resnet50 classification accuracy matrix by month and time of the day. (**D**) Matrix showing test images count for each month used in Resnet50 model.

confirm that these covariates are significant confounding variables. By explicitly balancing the dataset across these factors, we have ensured that the model is trained on a representative and challenging distribution of data, preventing it from learning “shortcut” solutions (e.g., associating a species with only one lighting condition or device type). The high, though imperfect, accuracy indicates that the model is learning true morphological features while also revealing the inherent difficulty of classification under challenging environmental conditions, such as low evening light.

Challenges in Bark Classification and Dataset Complexity

Deep learning models, particularly convolutional neural networks (CNNs)[15], are well-suited for tasks requiring intricate pattern recognition. However, our AI experiments on the BarkVisionAI dataset demonstrate that tree bark classification remains a challenging task, even for state-of-the-art deep learning architectures. While models such as ResNet50[16] achieved high classification accuracy (87.42%) on the dataset, a closer analysis highlights the underlying complexities and the necessity for further research to fine-tune models for improved reliability[17–19].

The variation in results across different architectures points to a broader challenge: extracting and distinguishing fine-grained patterns in bark textures is not trivial. Unlike standard object classification tasks, where shape and distinct features dominate, bark classification demands models that can capture minute textural and structural variations while remaining robust to environmental factors such as lighting, moisture, and seasonal changes (Fig. 7).

Overall, it can be said that the BarkVisionAI dataset’s design – encompassing images from different seasons, tree ages, and device types, forest types – ensures that these real-world complexities are well represented. This is a significant contribution, as it demonstrates that bark classification is a non-trivial problem requiring further advancements in deep learning techniques. Our results suggest that future work should explore:

- More sophisticated feature extraction methods to capture the nuanced texture differences in bark images.
- Fine-tuning of hyperparameters and domain-specific augmentations tailored to bark variability.

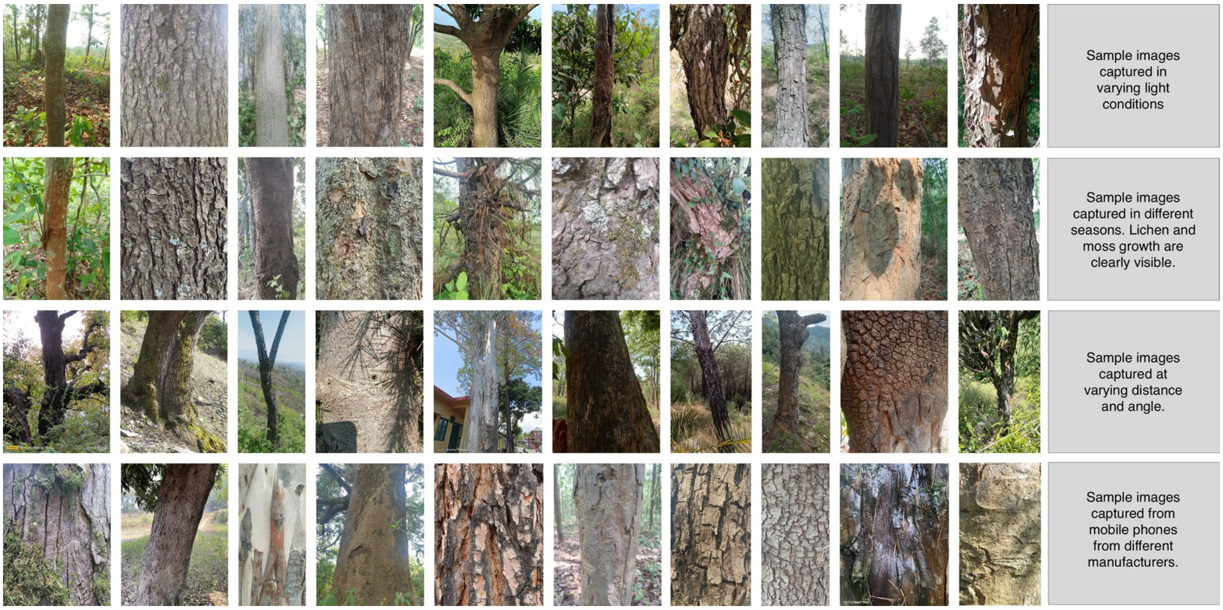

Sample images captured in varying light conditions
Sample images captured in different seasons. Lichen and moss growth are clearly visible.
Sample images captured at varying distance and angle.
Sample images captured from mobile phones from different manufacturers.

Fig. 7 Sample tree bark images illustrating variability across multiple parameters—such as lighting conditions (time of day), seasonal changes, shadows, external growth artifacts like moss, lichens, and other symbiotic or parasitic vegetation, as well as variations in distance, camera angle, and device type. This diversity introduces significant complexity to the dataset, making the classification task more challenging.

In summary, while the dataset provides a high-level encapsulation of the real-world variability and is aimed to enable high-accuracy classification with standard deep learning models, it also reveals the inherent challenges in tree species identification through bark images alone. This highlights the need for continued research to develop models that can more effectively learn and generalize across the natural variability present in tree bark textures.

Data availability

The dataset is available on Figshare: https://doi.org/10.6084/m9.figshare.28427246.

Code availability

The code to reproduce the model results are available on: GitHub.

Received: 21 February 2025; Accepted: 26 January 2026;
Published online: 03 February 2026

References

1. Baeten, L. *et al*. Identifying the tree species compositions that maximize ecosystem functioning in European forests. *Journal of Applied Ecology* **56**, 733–744 (2019).
2. Kim, T. K. *et al*. Identifying and extracting bark key features of 42 tree species using convolutional neural networks and class activation mapping. *Scientific Reports 2022 12:1* **12**, 1–13 (2022).
3. Fekri-Ershad, S. Bark texture classification using improved local ternary patterns and multilayer neural network. *Expert Syst Appl* **158** (2020).
4. Šulc, M. & Matas, J. Fine-grained recognition of plants from images. *Plant Methods* **13** (2017).
5. Ratajczak, R., Bertrand, S., Crispim-Junior, C. & Tougne, L. Efficient bark recognition in the wild. *VISIGRAPP 2019 - Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications* **4**, 240–248 (2019).
6. Carpentier, M., Giguere, P. & Gaudreault, J. Tree Species Identification from Bark Images Using Convolutional Neural Networks. *IEEE International Conference on Intelligent Robots and Systems* 1075–1081, https://doi.org/10.1109/IROS.2018.8593514 (2018).
7. Olson, D., Dinerstein, E., E. W.- & 2001, undefined. Terrestrial Ecoregions of the World: A New Map of Life on Earth: A new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. *academic.oup.comDM Olson, E Dinerstein, ED Wikramanayake, ND Burgess, GVN Powell, EC UnderwoodBioScience, 2001·academic.oup.com*, https://academic.oup.com/bioscience/article-abstract/51/11/933/227116 (2001).
8. Zheru, C., Li, H. & Wang, C. Plant species recognition based on bark patterns using novel Gabor filter banks. *Proceedings of 2003 International Conference on Neural Networks and Signal Processing, ICNNSP'03* **2**, 1035–1038 (2003).
9. Wu, F., Gazo, R., Benes, B. & Haviarova, E. Deep BarkID: a portable tree bark identification system by knowledge distillation. *Eur J For Res* **140**, 1391–1399 (2021).
10. Dombrowski, M., Prenner, A. & Kainz, B. Bias Assessment and Data Drift Detection in Medical Image Analysis: A Survey. https://arxiv.org/pdf/2409.17800v1 (2025).
11. Dallas, T. A. & Ten Caten, C. Linking geographic distribution and niche through estimation of niche density. *J Anim Ecol* **94**, 1221 (2025).
12. Geirhos, R. *et al*. Shortcut learning in deep neural networks. *Nature Machine Intelligence 2020 2:11* **2**, 665–673 (2020).
13. PlantCLEF 2024 | ImageCLEF/LifeCLEF - Multimedia Retrieval in CLEF. https://www.imageclef.org/node/315.

14. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* **2016-December**, 770–778 (2015).
15. Misra, D., Crispim-Junior, C. & Tougne, L. Patch-Based CNN Evaluation for Bark Classification. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **12540 LNCS**, 197–212 (2020).
16. Tan, M. & Le, Q. V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *36th International Conference on Machine Learning, ICML 2019* **2019-June**, 10691–10700 (2019).
17. Boudra, S., Yahiaoui, I. & Behloul, A. Plant identification from bark: A texture description based on Statistical Macro Binary Pattern. *Proceedings - International Conference on Pattern Recognition* **2018-August**, 1530–1535 (2018).
18. Wan, Y. Y. *et al*. Bark texture feature extraction based on statistical texture analysis. *2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, ISIMP 2004* 482–485, https://doi.org/10.1109/ISIMP.2004.1434106 (2004).
19. Robert, M., Dallaire, P. & Giguere, P. Tree bark re-identification using a deep-learning feature descriptor. *Proceedings - 2020 17th Conference on Computer and Robot Vision, CRV 2020* 25–32, https://doi.org/10.1109/CRV50864.2020.00012 (2020).
20. Chhatre, A. *et al*. BarkVisionAI: Novel dataset for rapid tree species identification. *Figshare*. https://doi.org/10.6084/m9.figshare.28427246 (2026).

Acknowledgements

We sincerely acknowledge the Forest Guards, Range Officers, and Divisional Forest Officers of the Himachal Pradesh Forest Department for their crucial support in data collection. We extend our gratitude to our research team and field experts for their dedicated efforts in conducting field surveys and ensuring data accuracy. We gratefully acknowledge the International Land and Forest Tenure Facility, Sweden, for providing essential funding support that enabled the design, field implementation, and curation of the BarkVisionAI dataset. Their commitment to advancing community-centered forest governance and evidence-based research made it possible to undertake large-scale data collection across diverse forest landscapes in India. This support was instrumental in establishing the technical and organizational foundation required to generate an open, high-quality dataset for the global scientific community.

Author contributions

A.C.: supervision, conceptualization, writing—review and editing, and funding acquisition. N.S.: methodology, data acquisition, and formal analysis. A.K.P.: conceptualization, methodology, data acquisition, analysis, and writing—original draft. P.R.: data acquisition, and writing—review and editing. M.J.: methodology, writing—review and editing.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.