Heliyon 10 (2024) e40606

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

IndoHerb: Indonesia medicinal plants recognition using transfer learning and deep learning

Muhammad Salman Ikrar Musyaffa [a], Novanto Yudistira [a,*], Muhammad Arif Rahman [a], Ahmad Hoirul Basori [c], Andi Besse Firdausiah Mansur [c], Jati Batoro [b]

[a] Informatics Engineering, Faculty of Computer Science, Brawijaya University, Malang, 65145, East Java, Indonesia
[b] Departement of Biology, Faculty of Mathematics and Natural Science, Brawijaya University, Malang, 65145, East Java, Indonesia
[c] Faculty of Computing and Information Technology in Rabigh, King Abdulaziz University, Rabigh, 21911, Saudi Arabia

ARTICLE INFO

Keywords:
Transfer learning
Convolutional neural network
Computer vision
Medicinal plant
Images recognition

ABSTRACT

The rich diversity of herbal plants in Indonesia holds immense potential as alternative resources for traditional healing and ethnobotanical practices. However, the dwindling recog-nition of herbal plants due to modernization poses a significant challenge in preserving this valuable heritage. The accurate identification of these plants is crucial for the continuity of traditional practices and the utilization of their nutritional benefits. Nevertheless, the manual identification of herbal plants remains a time-consuming task, demanding expert knowledge and meticulous examination of plant characteristics. In response, the application of computer vision emerges as a promising solution to facilitate the efficient identification of herbal plants. This research addresses the task of classifying Indonesian herbal plants through the implementation of transfer learning of Convolutional Neural Networks (CNN). To support our study, we curated an extensive dataset of herbal plant images from Indonesia with careful manual selection. Subsequently, we conducted rigorous data preprocessing, and classification utilizing transfer learning methodologies with five distinct models: ResNet, DenseNet, VGG, ConvNeXt, and Swin Transformer. Our comprehensive analysis revealed that ConvNeXt achieved the highest accuracy, standing at an impressive 92.5 %. Additionally, we conducted testing using a scratch model, resulting in an accuracy of 53.9 %. The experimental setup featured essential hyperparameters, including the ExponentialLR scheduler with a gamma value of 0.9, a learning rate of 0.001, the Cross-Entropy Loss function, the Adam optimizer, and a training epoch count of 50. This study's outcomes offer valuable insights and practical implications for the automated identification of Indonesian medicinal plants, contributing not only to the preservation of ethnobotanical knowledge but also to the enhancement of agricultural practices through the cultivation of these valuable resources. The Indonesia Medicinal Plant Dataset utilized in this research is openly accessible at the following link: https://github.com/Salmanim20/indomedicinalplant.

1. Introduction

Indonesia, with its tropical climate and abundant biodiversity, is home to a vast array of plant species, among which herbal plants

* Corresponding author.
E-mail address: yudistira@ub.ac.id (N. Yudistira).

https://doi.org/10.1016/j.heliyon.2024.e40606
Received 14 November 2024; Received in revised form 19 November 2024; Accepted 20 November 2024
Available online 21 November 2024

hold a prominent position. This archipelagic nation boasts approximately 30,000 distinct plant species, with around 9600 classified as herbal plants. These herbal plants have been used for generations as a means of traditional healing. However, with ongoing societal and technological advancements, the use of herbal remedies has gradually declined, as modern pharmaceuticals increasingly supplant traditional practices. This transition, while facilitating access to advanced medical care, threatens to overshadow the long-established efficacy of herbal treatments. Therefore, it is crucial to explore innovative approaches that revive the recognition and use of herbal plants, especially in regions where economic or geographical constraints limit access to modern medicine.

The significance of herbal plants extends beyond their therapeutic potential. In an era marked by heightened health awareness and growing demand for natural, chemical-free food products, herbal plants have found renewed importance. They play a pivotal role in producing organic, health-conscious foods and are central to a natural, holistic lifestyle. This resurgence of interest in herbal plants has led to a surge in organic cultivation practices, aligning with the broader trend of environmental sustainability [1].

Indonesia's rich diversity of herbal plants holds immense potential as alternative resources for traditional healing and ethnobotanical practices. Historically, these plants have been integral to Indonesian culture and medicine, serving as natural remedies for various ailments and forming a core part of the country's cultural heritage [2–4]. Herbal plants such as ginger, turmeric, and betel leaf have been used in traditional Jamu (herbal medicine) for centuries, demonstrating their longstanding significance in maintaining health and wellness [5].

However, the dwindling recognition and use of herbal plants due to modernization poses a significant challenge in preserving this valuable heritage. The rapid urbanization and industrialization in Indonesia have led to the erosion of traditional knowledge, with younger generations often favoring modern pharmaceuticals over traditional remedies. Additionally, deforestation and habitat destruction threaten the natural habitats of many of these plants, further endangering their availability and continuity [6].

Accurate identification of herbal plants is essential, but it presents a formidable challenge, often demanding an exhaustive understanding of plant phenotypes and intricate botanical knowledge. Plant identification encompasses a multitude of criteria, including color, flower morphology, leaf structure, texture, and overall plant architecture. The vast number of plant species, coupled with morphological similarities among closely related species, compounds this challenge [7]. The advent of computer vision technology in recent years offers a promising solution to expedite and enhance the accuracy of plant identification. Utilizing integrated cameras and machine learning algorithms, individuals can access plant information swiftly and accurately. Consequently, computer vision has become a focal point of research in the domain of herbal plant identification, leveraging various plant features such as leaves, roots, and fruits [8].

Previous research by Quoc and Hoang [8] explored herbal plant identification using the Scale-Invariant Feature Transform (SIFT) [9] and Speeded Up Robust Features (SURF) [10] algorithms. The study employed two resolution versions, 256×256 and 512×512, yielding accuracy rates ranging from 21 % to 37.4 %. Additionally, Liantoni [11] investigated classification methods, including Naive Bayes [12] and K-Nearest Neighbor [13], achieving accuracy rates of 70.83 %–75 %. While these results provide valuable insights, they underscore the need for further improvement in classification accuracy.

This study aims to investigate the potential of using transfer learning with Convolutional Neural Networks (CNNs) in identifying Indonesian herbal plants within a curated IndoHerb database. Transfer learning offers a compelling advantage by combining feature extraction and classification algorithms, optimizing learning efficiency. Its proficiency in object classification within images makes it a promising candidate for enhancing prediction accuracy. Additionally, this study seeks to augment the research landscape by assembling a dedicated dataset of Indonesian herbal plants, meticulously collected through Google Images searches. However, it is important to acknowledge the limitations of this research. These limitations include the utilization of only five pre-trained models: ResNet, DenseNet, VGG11, ConvNeXt, and Swin Transformer. Additionally, the study focused exclusively on 100 classes, selected based on information obtained from an online source. Furthermore, the Dynamic Learning Rate scheduler used was restricted to the ExponentialLR scheduler.

In conclusion, our contribution can be summarized into three folds.

1. Investigating the application of transfer learning with CNNs for the identification of Indonesian herbal plants within the IndoHerb database.
2. The creation of a dedicated dataset of Indonesian herbal plants, which is a valuable resource for future studies in the field of herbal plant identification.
3. Advancement of herbal plant identification through the application of cutting-edge technology, fostering a renewed appreciation for the invaluable wealth of botanical resources in Indonesia and beyond.

2. Related works

Prior research in the field of medicinal plant classification has laid a significant foundation for our current study. Notable contributions include the work of Quoc & Hoang [8], who conducted extensive research on Vietnam Medicinal Plants. Their study employed the Scale-Invariant Feature Transform (SIFT) [9] and Speeded Up Robust Features (SURF) [10] algorithms and featured a dataset comprising 20,000 herbal plant images. The research meticulously explored two resolution versions, 256×256 and 512×512, with varying accuracy rates. Specifically, in the 256×256 version, the SURF method achieved an accuracy of 21 %, while the SIFT method reached 28 %. In the higher-resolution 512×512 version, the SURF method improved to an accuracy of 34.7 %, and the SIFT method achieved 37.4 %.

Liantoni [11] delved into herbal leaf classification, utilizing the Naive Bayes [12] and K-Nearest Neighbor [13] methods. Their dataset comprised 120 images, with 96 images allocated for training and 24 for testing. The study revealed the superiority of the Naive

Bayes method, achieving an accuracy rate of 75 % compared to the K-Nearest Neighbor method's 70.83 %. Both methods underwent rigorous testing, spanning 100 epochs.

Naeem et al. [14] contributed to the field with their focus on the classification of medicinal plant leaves. Their dataset included 6000 leaf images distributed across six classes, with each class containing 1000 images. The study employed an array of classification algorithms, such as Multi-Layer Perceptron (MLP) [15], Logit-Boost (LB) [16], Bagging (B) [17], Random Forest (RF) [18], and Simple Logistic (SL) [19]. Two image sizes, 220 × 220 and 280 × 280, were considered, resulting in accuracy rates ranging from 92.56 % to 99.01 %. Notably, the Multi-Layer Perceptron method outperformed other methods, asserting its effectiveness in medicinal plant leaf classification.

Recent advancements in deep learning have significantly enhanced the classification of medicinal plant images through various techniques, including transfer learning, fine-tuning [20,21], and ensemble models [22]. Transfer learning, where pre-trained models are adapted to specific tasks, has proven effective in addressing the challenges associated with limited datasets. By fine-tuning these pre-trained models, researchers can adjust them to better recognize features unique to medicinal plants, thereby improving

Table 1
List of plant species along with the number of samples before augmentation for classification.

Plant Species	Plant Family	No. of Samples	Plant Species	Plant Family	No. of Samples	Plant Species	Plant Family	No. of Samples
Abelmoschus esculentus	*Malvaceae*	85	*Acorus calamus*	*Acoraceae*	100	*Aloe vera*	*Asphodelaceae*	45
Alstonia scholaris	*Apocynaceae*	75	*Amaranthus spinosus*	*Amaranthaceae*	48	*Andrographis paniculata*	*Acanthaceae*	87
Annona muricata	*Annonaceae*	80	*Annona squamosa*	*Annonaceae*	57	*Anredera cordifolia*	*Basellaceae*	100
Apium graveolens	*Apiaceae*	70	*Artocarpus heterophyllus*	*Moraceae*	90	*Artocarpus integer*	*Moraceae*	82
Averrhoa bilimbi	*Oxalidaceae*	80	*Blumea balsamifera*	*Asteraceae*	88	*Borreria hispida*	*Rubiaceae*	30
Caesalpinia sappan	*Fabaceae*	30	*Caladium cicolor*	*Araceae*	81	*Calendula officinalis*	*Asteraceae*	78
Canangium odoratum	*Annonaceae*	78	*Catharanthus roseus*	*Apocynaceae*	70	*Celosia cristata*	*Amaranthaceae*	83
Centella asiatica	*Apiaceae*	75	*Cestrum nocturnum*	*Solanaceae*	50	*Citrus hystrix*	*Rutaceae*	83
Clinalanthus nutans	*Acanthaceae*	100	*Clitoria ternatea*	*Fabaceae*	75	*Crinum asiaticum*	*Amaryllidaceae*	92
Curcuma domestica	*Zingiberaceae*	44	*Cyclea barbata*	*Menispermaceae*	100	*Cymbopogon nardus*	*Poaceae*	84
Derris elliptica	*Fabaceae*	27	*Desmodium triquitrum*	*Fabaceae*	78	*Dioscorea hispida*	*Dioscoreaceae*	92
Eleutherine americana	*Iridaceae*	100	*Euodia suaveolens*	*Rutaceae*	46	*Eupatorium triplinerve*	*Asteraceae*	73
Euphorbia tirucalli	*Euphorbiaceae*	87	*Euphoria longan*	*Sapindaceae*	73	*Ficus carica*	*Moraceae*	46
Ficus septica	*Moraceae*	100	*Graptophyllum pictum*	*Acanthaceae*	100	*Gynura segetum*	*Asteraceae*	100
Hibiscus rosa-sinensis	*Malvaceae*	75	*Hibiscus sabdariffa*	*Malvaceae*	82	*Houttuynia cordata*	*Saururaceae*	100
Hydrocotyle sibthorpioides	*Araliaceae*	89	*Impatiens balsamina*	*Balsaminaceae*	83	*Isotoma longiflora*	*Campanulaceae*	80
Jasminum sambac	*Oleaceae*	58	*Jatropha multifida*	*Euphorbiaceae*	100	*Kaempferia galanga*	*Zingiberaceae*	28
Melaleuca leucadendra	*Myrtaceae*	79	*Melia azedarach*	*Meliaceae*	31	*Melissa officinalis*	*Lamiaceae*	75
Merremia mammosa	*Convolvulaceae*	100	*Michelia alba*	*Magnoliaceae*	83	*Mirabilis jalapa*	*Nyctaginaceae*	100
Morinda citrifolia	*Rubiaceae*	50	*Morus alba*	*Moraceae*	79	*Muraya paniculata*	*Rutaceae*	83
Murraya koenigii	*Rutaceae*	100	*Nepeta cataria*	*Lamiaceae*	100	*Nothopanax scutellarium*	*Araliaceae*	34
Ocimum americanum	*Lamiaceae*	32	*Ocimum basilicum*	*Lamiaceae*	82	*Olea europaea*	*Oleaceae*	65
Orthosiphon spicatus	*Lamiaceae*	89	*Pandanus amaryllifolius*	*Pandanaceae*	83	*Parameria laevigata*	*Apocynaceae*	20
Peperomia pellucida	*Piperaceae*	76	*Phaleria macrocarpa*	*Thymelaeaceae*	47	*Physalis angulata*	*Solanaceae*	78
Phytolacca americana	*Phytolaccaceae*	75	*Piper betle*	*Piperaceae*	81	*Piper sarmentosum*	*Piperaceae*	100
Pluchea indica	*Asteraceae*	100	*Polianthes tuberosa*	*Asparagaceae*	74	*Polyscias scutellaria*	*Araliaceae*	92

classification accuracy.

However, while ensemble methods can provide robust predictions by combining multiple models, they can also introduce a high computational burden due to the large number of parameters involved. This can make ensemble approaches less effective in scenarios where computational resources are limited or when dealing with very large datasets. Despite this, transfer learning combined with fine-tuning remains a powerful strategy for optimizing model performance on specialized medicinal plant datasets, balancing accuracy and efficiency without the excessive complexity of ensemble models.

Drawing inspiration from these notable research contributions, our current study aims to extend the existing body of knowledge by introducing a new dataset of Indonesian medicinal plants. This dataset, curated independently using the Google Images search engine, forms the basis for our exploration of transfer learning methods from convolutional neural networks to enhance classification accuracy. The challenges and strategies encountered during dataset collection and preparation inform the robustness of our forthcoming research.

In summary, the related works discussed here provide valuable insights and benchmarks in medicinal plant classification, setting the stage for our contributions to this evolving field through the creation of a new dataset and the application of advanced machine learning techniques.

3. IndoHerb dataset

Based on the previous research, this study aims to create a new dataset of medicinal plants from Indonesia, collected independently through the Google Images search engine called IndoHerb. To ensure that the plant images correspond to the classification, we cross-referenced the medicinal plants with the list provided by the Medicinal Plant Maintenance Installation at the Baturaja Health Research and Development Center, Indonesian Ministry of Health [23]. Subsequently, the images were systematically collected from the web. Finally, the dataset underwent validation and meticulous selection by a professor of biology—an expert in Plant Taxonomy and Ethnobotany—to authenticate the collected images of Indonesian medicinal plants. The datasets is tested using the transfer learning method from the convolutional neural network.

During the dataset collection process, the initial step involved determining which species of medicinal plants from Indonesia would be utilized for research. To identify these species, searches were conducted on various websites that provided the names of medicinal plant species from Indonesia, resulting in the collection of 100 species for research. Following this, an image search was conducted for each species of medicinal plants using the Google Images search engine.

From the search results, images that met the criteria for each species were manually selected from the top search results. This manual selection was necessary because, at times, there were discrepancies, such as instances where identical images appeared from different sources. While this issue was minimized in the study, inadvertent selections of identical images were left unchanged. Additionally, some images were excluded due to quality issues, such as scribbles or watermarks. Another challenge involved less-known or rarely found species, resulting in a limited number of search results (e.g., approximately 30 images meeting the criteria). To address this, an image augmentation process was implemented on the obtained images, including horizontal flips, vertical flips, and rotations. Ultimately, all images were resized to 128×128. The final dataset of medicinal plants from Indonesia comprises 10,000 images across 100 species, with each species containing 100 images.

Based on Table 1, it can be seen that there are various numbers of images collected from each species before the augmentation process is carried out. This is due to the limited number of images that can be selected for research due to several problem factors as previously described. The species used in the dataset is shown in Table 1.

The number of Indonesian Medicinal Plant datasets that have been collected is 7391 images. For this reason, an augmentation process is carried out for each species that do not meet the target number of 100 so that each species has the same number. In the end, we get 10,000 images in total where the 2609 images result from the augmentation process in the form of horizontal flip, vertical flip, and rotation.

Indonesian Medicinal Plant Database (IndoHerb) and the Vietnam Medicinal Plant Database reveal notable similarities and differences that influence their respective contributions to medicinal plant classification research. Both datasets primarily consist of images captured in natural environments, which introduces variability in illumination, scale, and background complexity. This variability reflects real-world conditions where plants are found in diverse settings, making both datasets valuable for training models to handle a wide range of natural scenarios. They include images of leaves, flowers, and whole plants, often taken in natural settings with varying backgrounds. This introduces additional complexity, as models must contend with environmental factors and lighting variations representative of real-world conditions.

Both datasets are curated to ensure high-quality images, though their approach differs.

IndoHerb excludes images with watermarks, scribbles, or low resolution, and employs augmentation techniques to balance the dataset across species. In contrast, the Vietnam dataset, with its larger volume of 20,000 images across 200 species, encompasses a wider range of image qualities, reflecting natural imperfections. Together, these datasets enhance research by ensuring that models are not only accurate but also resilient and adaptable to varied and realistic conditions. This combined approach strengthens the overall research efforts and contributes to the development of more reliable medicinal plant classification systems.

4. Methodology

4.1. Gathering data

The collected data consists of images of herbal plants sourced from two main datasets: the Vietnam Medicinal Plant and the Indonesia Medicinal Plant dataset. The Vietnam dataset comprises a total of 20,000 images categorized into 200 species. On the other hand, the Indonesia Medicinal Plant dataset was obtained independently through Google Images search engine and includes 10,000 images from 100 species. Example images from the Indonesia Medicinal Plant Dataset are illustrated in Fig. 1.

Based on Fig. 1, it showcases examples of 5 species from the Indonesian herbal plant dataset, each consisting of 5 image examples. These images were collected independently, representing species such as *Abelmoschus esculentus*, *Acorus calamus*, *Aloe vera*, *Alstonia scholaris*, and *Amaranthus spinosus*. Since the images were sourced from the Google Images search engine, the selected images exhibit a diverse range within each species.

4.2. Preprocessing data

The collected data undergoes preprocessing. Initially, the dataset is read, followed by subsequent transformations such as resizing, rotating, and normalizing the image data. Images are resized to dimensions of 128 × 128 and their pixel values are normalized to fall within the range of 0–1. Following preprocessing, the data is split into training and testing sets.

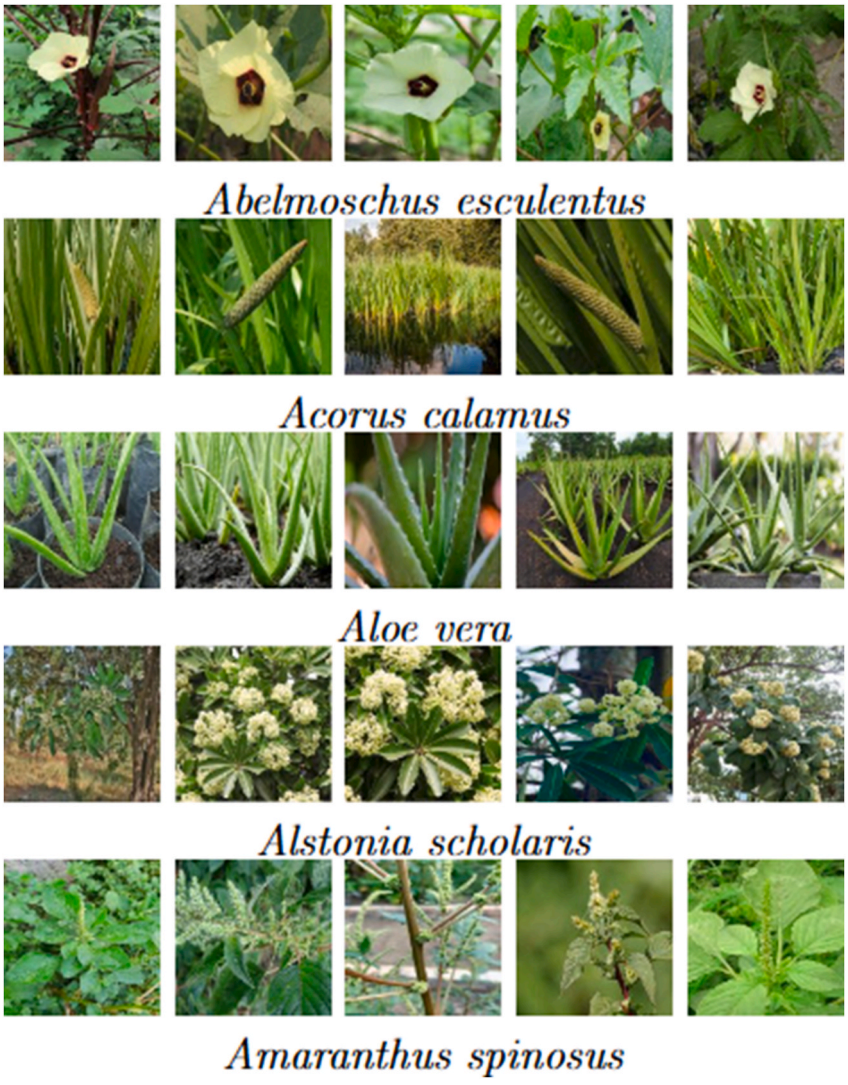

Fig. 1. Example images from five selected species of Indonesia medicinal plant dataset.

4.3. Structuring model

In this study, the Transfer Learning method is employed for training. The Transfer Learning models utilized include pre-trained ResNet34 [24], DenseNet121 [25], VGG11 bn [26], ConvNeXt [27], and Swin Transformer [28] models. Additionally, tests are conducted using the scratch model. The code is prepared to initiate training using Transfer Learning with these models. The model's capacity to discriminate across plant species is measured by its accuracy in classifying each species into its corresponding class.

The architecture of the scratch model begins with an initial convolutional layer that uses 32 output channels and a 3×3 kernel, followed by a 2×2 Max Pooling layer. The second convolutional layer utilizes 64 output channels with a 3×3 kernel, followed by another 2×2 Max Pooling layer. The third convolutional layer employs 128 output channels with a 3×3 kernel, again followed by a 2×2 Max Pooling layer. After the convolutional layers, the output is flattened, transforming the matrix into a vector, which is then passed through a linear classification layer with a ReLU activation function. The final output layer classifies the input into one of 100 classes.

Transfer learning played a pivotal role in our methodology, significantly enhancing the performance of models on the Indonesian Medicinal Plant dataset. By leveraging models pre-trained on extensive datasets like ImageNet, we capitalized on their well-developed feature extraction capabilities. These models, originally trained on over a million images spanning a thousand classes, possess a deep understanding of diverse visual features, including object recognition, texture analysis, and edge detection.

Despite their strong generalization abilities, the task of classifying Indonesian medicinal plants required more specific knowledge. To address this, we employed fine-tuning, a process that adapts the pre-trained models to our dataset's unique characteristics. This involved re-training all layers of the models, enabling them to recalibrate their parameters based on the distinct visual patterns in the medicinal plant images. By fine-tuning all layers, the models were able to adjust both high-level features, such as shapes and objects, and low-level features, like edges and textures, to better suit our dataset.

Moreover, fine-tuning, rather than training from scratch, considerably reduced the computational cost and time needed to achieve high performance. The models could swiftly con-verge to an optimal solution by refining the already-learned features from the pre-training phase, avoiding the need to learn from a randomly initialized state. This approach was particularly effective given the relatively small size of the Indonesian Medicinal Plant dataset compared to datasets like ImageNet.

As a result, the fine-tuned models exhibited an enhanced ability to recognize and differentiate between various medicinal plant species, many of which have subtle visual distinctions. This strategy improved classification accuracy and established a robust framework that could be adapted for other specialized plant datasets.

Table 2 shows a variety of deep learning models were selected for their distinct advantages in image classification tasks. ResNet (Residual Networks) was chosen for its ability to train very deep networks efficiently by using residual connections, addressing the vanishing gradient problem, and delivering state-of-the-art performance. DenseNet (Dense Convolutional Network) was included for its innovative architecture that connects each layer to every other layer, improving information flow and parameter usage efficiency while mitigating the vanishing gradient issue. The VGG (Visual Geometry Group) model, known for its straight-forward and uniform architecture with small convolutional filters, was selected for its reliable baseline performance in image classification. ConvNeXt represents a modern architectural approach that blends the strengths of convolutional neural networks (CNNs) and transform-ers, offering robust performance by integrating features from both types of models. Finally, the Swin Transformer, a vision transformer model, was chosen for its ability to capture long-range dependencies through self-attention mechanisms, making it particularly effective for handling complex image classification tasks. Together, these models provide a comprehensive comparison across different architectural paradigms, enhancing the study's exploration of classification performance on the Indonesian Medicinal Plant dataset.

4.4. ConvNeXt

ConvNeXt represents a significant evolution in the design of convolutional neural networks (ConvNets), setting itself apart from

Table 2
Rationale for selecting specific models.

Model	Rationale	Reference
ResNet	• Addresses vanishing gradient problem •Enables deep network training with residual connections •State-of-the-art for various classifica- tion tasks	(He et al., 2016) [24]
DenseNet	•Connects each layer to every other layer •Improves parameter efficiency •Reduces vanishing gradient problem	(Huang et al., 2017) [25]
VGG	•Uses uniform small convolutional filters •Simple, effective baseline for classification •Consistent performance across benchmarks	(Simonyan and Zisser- man, 2015) [26]
ConvNeXt	•Combines CNN and transformer elements •Efficient feature extraction with attention mechanisms •Modern, robust performance for image classification	(Liu et al., 2022) [27]
Swin Trans-former	•Captures long-range dependencies with self-attention •Excels in complex classification tasks •Models global context effectively	(Liu et al., 2021) [28]

both traditional ConvNets and newer models like Vision Transformers (ViTs) and Swin Transformers. The core difference lies in ConvNeXt's approach to modernizing the ConvNet architecture to compete directly with the performance of state-of-the-art Transformer models while retaining the fundamental principles of convolution.

Vision Transformers introduced a paradigm shift in visual recognition by utilizing self-attention mechanisms instead of convolutions, allowing for more flexibility in processing visual data. However, ViTs lack the inductive biases inherent in ConvNets, such as translation invariance and locality, which makes them less efficient in certain tasks like object detection and semantic segmentation. To address this, hybrid models like Swin Transformers reintroduced ConvNet-like hierarchical features into the Transformer architecture, improving their practicality for a broader range of computer vision tasks. Despite this, the success of these hybrids is often attributed more to the transformative capabilities of self-attention than to the convolutional components.

ConvNeXt, on the other hand, focuses on refining and modernizing the ConvNet architecture itself. Instead of adopting self-attention, ConvNeXt incorporates design elements inspired by Transformers, such as larger kernel sizes, fewer non-linearities, and improved normalization techniques. By doing so, ConvNeXt maintains the simplicity and efficiency of ConvNets while achieving performance levels comparable to top-tier Transformers. This architectural evolution allows ConvNeXt to excel in tasks like ImageNet classification, COCO detection, and ADE20K segmentation, demonstrating that ConvNets, when thoughtfully updated, can still serve as powerful, general-purpose vision backbones.

In essence, ConvNeXt bridges the gap between traditional ConvNets and Transformer-based models, offering a modernized convolutional architecture that competes effectively with the latest advances in visual recognition while preserving the inherent strengths of convolutions.

4.5. Model's capacity to discriminate across plant families and species

The classification of medicinal plants, particularly across diverse families and species, is inherently challenging due to the morphological similarities among closely related species and the wide range of phenotypic traits within and across plant families. This task requires models that can effectively capture and discriminate subtle differences in plant structures, such as leaf shape, venation patterns, flower morphology, and overall plant architecture.

The adoption of transfer learning using CNNs has shown promise in enhancing the model's capacity to discriminate across plant families and species. Transfer learning allows models to leverage pre-trained weights from large-scale datasets, which have already captured a vast array of visual features. When fine-tuned on a specific dataset like IndoHerb, these models can adapt to recognize the unique features of Indonesian medicinal plants, even with a relatively small number of training images per class.

The pre-trained models used in this research, including ResNet, DenseNet, VGG11, ConvNeXt, and Swin Transformer, are well-suited for handling the complexity of plant image classification. These models excel in feature extraction, enabling them to discern fine-grained details that differentiate one species from another, even within the same family. The hierarchical structure of CNNs, which captures low-level features (e.g., edges, textures) in the initial layers and more complex patterns in deeper layers, is particularly effective in distinguishing between species with subtle morphological differences.

At the family level, the model's capacity to discriminate is supported by the distinct visual features that are often characteristic of plant families. For example, the model can identify the unique leaf venation patterns typical of the Asteraceae family or the specific flower structures associated with the Fabaceae family. By learning these family-specific features, the model can accurately classify plants at the family level, even when faced with variations in lighting, angle, or image quality.

At the species level, the model faces a more nuanced challenge, as species within the same family often share many visual characteristics. Here, the model's ability to focus on subtle differences—such as variations in leaf margins, flower color, or growth habit—becomes critical. The IndoHerb dataset's diverse range of species, each represented by multiple images, allows the model to learn these fine-grained distinctions. This capability is further enhanced by data augmentation, which introduces variability in the training set, encouraging the model to generalize well to unseen images.

4.6. Testing model

Testing is conducted to evaluate the CNN model designed. The testing comprises both training and testing stages. During the training stage, the CNN models are evaluated using previously prepared training data. A total of 20,000 and 10,000 images from two different datasets are utilized. These datasets are split into 60 % training data and 40 % testing data. After the completion of the training process, testing is commenced. For this study, 8000 and 4000 images are used for testing, with 40 images allocated for each species of herbal plant. It's crucial to note that the testing stage employs different images from those used in the training process to accurately assess the performance of the model.

4.7. Evaluation metrics

Accuracy is a fundamental metric used to evaluate the performance of a classification model. It measures the proportion of correctly predicted instances (both positive and negative) out of the total number of cases. The formula for accuracy is given by:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

where TP (True Positives) represents the number of correctly predicted positive cases, TN (True Negatives) denotes the number of correctly predicted negative cases, FP (False Positives) indicates the number of negative cases incorrectly classified as positive, and FN (False Negatives) refers to the number of positive cases incorrectly classified as negative. Accuracy is useful for providing a general measure of a model's performance. However, it may not always be reliable, especially in scenarios where there is an imbalance in the class distribution, as it might not reflect the true performance across all classes.

Precision focuses on the quality of the positive predictions made by the model. It is defined as the ratio of true positive predictions to the sum of true positive and false positive predictions. The formula for precision is

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

where TP is the number of true positives and FP is the number of false positives. Precision is critical in situations where the cost of false positives is high, as it measures how many of the predicted positives are actually correct. For example, in medical diagnostics, high precision ensures that a large proportion of positive test results are accurate, reducing the risk of false alarms.

Recall, also known as sensitivity or true positive rate, measures a model's ability to identify all relevant positive cases. It is the ratio of true positive predictions to the total number of actual positives. The recall is calculated using the formula:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

where TP represents the true positives and FN stands for false negatives. Recall is especially important when missing a positive case has severe consequences. For instance, in disease screening, high recall ensures that most of the actual positive cases are identified, which is crucial for effective treatment.

The F1-Score is a comprehensive metric that combines both precision and recall into a single measure. It is the harmonic mean of precision and recall and is given by:

$$\text{F1} - \text{Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

where Precision and Recall are as defined above. The F1-Score is particularly useful when the balance between precision and recall is important, and it is often used in situations with imbalanced datasets where neither precision nor recall alone would provide a complete picture of model performance. By incorporating both metrics, the F1-Score offers a more holistic view of a model's effectiveness in handling both false positives and false negatives.

4.8. Differences among metrics

In the domain of multi-class herbal classification, the evaluation metrics of recall, precision, and accuracy are essential for assessing the performance of classification models, each offering distinct insights into the model's effectiveness across multiple classes.

Recall is particularly critical when the objective is to ensure comprehensive identification of herbs within each class. For instance, in a classification task involving various herb classes such as *Abelmoschus esculentus*, *Acorus calamus*, and *Aloe vera*, recall quantifies the proportion of actual instances of each herb that the model correctly classifies. High recall indicates that the model effectively captures most instances of a specific herb, minimizing the occurrence of false negatives. This metric is especially relevant in applications where the failure to identify a particular herb could have significant implications, such as in the field of herbal medicine, where missing a crucial herb could lead to incorrect treatment recommendations.

Precision, on the other hand, focuses on the accuracy of the model's predictions for each class. In the context of the same herb classes, precision measures the proportion of herbs predicted to belong to a particular class, such as *Abelmoschus esculentus*, that are indeed correctly classified. High precision is indispensable in scenarios where the consequences of misclassification are substantial, such as in culinary applications or pharmacology, where the incorrect identification of an herb could lead to adverse effects or incorrect usage.

Accuracy provides an aggregate measure of the model's overall performance across all classes. It represents the proportion of correctly classified instances relative to the total number of instances. While accuracy offers a broad overview of the model's efficacy, it may not adequately reflect the model's performance in multi-class scenarios, particularly when there is class imbalance. For example, if *Aloe vera* is overrepresented in the dataset, a model that predominantly classifies instances as *Aloe vera* might achieve high accuracy, yet perform poorly on less common classes like *Acorus calamus* or *Abelmoschus esculentus*. This can lead to a skewed perception of the model's true effectiveness.

5. Experiments

5.1. Testing the Vietnam Medicinal Plant Dataset

In this study, the dataset utilized is the Vietnam Medicinal Plant Dataset, which is tested on several models including pre-trained ResNet34 [24], DenseNet121 [25], VGG11 bn [26], ConvNeXt [27], Swin Transformer [28], and the Scratch model. During testing, a dynamic learning rate approach is employed. The scheduler used for this dynamic learning rate is the ExponentialLR [29] scheduler,

configured with a gamma value of 0.9. The learning rate for this test is set at 0.001. The dataset being tested consists of a total of 20,000 images, with a training-to-testing data distribution ratio of 60:40, resulting in 12,000 training images and 8000 testing images. There are 200 species in this dataset, with each species containing 60 training images and 40 testing images. The testing spans 50 epochs and utilizes the Adam optimizer [30], along with the Cross-Entropy Loss function [31].

Based on the information presented in Table 3, which summarizes the performance metrics of various models in our experiment, several key observations can be made regarding their accuracy and architecture. Firstly, when considering all six models in our study, ConvNeXt base emerges as the top-performing pre-trained model, boasting an impressive testing accuracy of 92.78 %. This indicates that ConvNeXt base excels in accurately classifying images and outperforms the other models under evaluation. Conversely, the model trained from scratch exhibits the lowest testing accuracy among the models, achieving a modest accuracy score of 48.49 %. This outcome can be attributed to the simplicity of the Scratch model's architecture when compared to the more complex, pre-trained models. The Scratch model's limitations in terms of learned features and representations likely contributed to its comparatively lower performance. While the accuracy of 65.04 % is respectable for Swin t, it lags behind the better-performing pre-trained models, especially when considering the differences in image resolution used. Moreover, when comparing the five pre-trained models—ResNet34, VGG11 bn, DenseNet121, ConvNeXt base, and Swin t—Swin t records the lowest accuracy. It is important to note that the evaluation is not solely based on accuracy but also on other metrics such as training and testing loss. These metrics provide deeper insights into how each model learned and improved during the training and testing phases. For a comprehensive understanding of the model's dynamics, including loss and accuracy graphs, we delve into a more detailed discussion in the subsequent sections of this paper.

Based on Table 4, the presented results offer a comprehensive evaluation of diverse neural network models in the context of a specific task, utilizing key performance metrics such as precision, recall, and F1-Score. The models under consideration encompass well-established architectures, including ResNet34, DenseNet121, VGG11 bn, ConvNeXt, and Swin t, along with a model trained from scratch denoted as "Scratch."

Given the reported metrics of Precision, Recall, and F1-Score, along with the corresponding model names and resolutions, the first model, ResNet34, with a resolution of 128^2, achieves a well-balanced performance, demonstrating competitive precision, recall, and F1-Score. The reported values suggest that ResNet34 is effective in identifying and classifying instances, with a slightly higher focus on precision.

Moving on to the second model, DenseNet121, also with a resolution of 128^2, stands out with high precision, recall, and F1-Score. These results indicate that DenseNet121 excels in capturing both positive and negative instances, with a particular emphasis on precision. The resolution of 128^2 signifies the input size used for this model in the evaluation.

The third model, VGG11 bn, at a resolution of 128^2, demonstrates notable precision but slightly lower recall and F1-Score. While its precision is high, the trade-off appears to be a compromise in recall. The values suggest that VGG11 bn is effective in minimizing false positives, but it may miss some instances, impacting its recall and overall F1-Score.

The fourth model, ConvNeXt, with a resolution of 128^2, outperforms others with impressive precision, recall, and F1-Score. These results indicate that ConvNeXt excels in both identifying and capturing instances with high accuracy, making it a robust choice for the given task.

However, the fifth model, Swin t, with a higher resolution of 224^2, exhibits lower precision, recall, and F1-Score values. This suggests that increasing the resolution may not have positively impacted the model's performance in this context, potentially indicating a trade-off between resolution and classification accuracy.

Finally, the sixth model, Scratch, at a resolution of 128^2, displays comparatively lower precision, recall, and F1-Score values. These results indicate that the Scratch model, developed without leveraging pre-trained weights, may not perform as well as the other models evaluated in this study.

The models have been ranked based on the normalized leverage factor, γ_ω, which reflects the relative importance or leverage of each model's precision compared to others [32]. **ConvNeXt** ranks highest with a γ_ω of 0.222, indicating that it has the most significant impact in terms of precision, recall, and F1-Score among the models evaluated. This suggests that ConvNeXt not only has strong performance metrics but also has the most balanced and robust leverage across these metrics. **VGG11 bn** and **DenseNet121** follow closely behind, with slightly lower but still competitive γ_ω values, indicating solid performance with good precision and recall

Table 3
Training and testing results of Vietnam medicinal plant dataset.

No.	Model Name	Resolution	Training		Testing	
			Loss	Accuracy	Loss	Accuracy
1.	SIFT [8]	512^2	–	–	–	0.374
2.	SURF [8]	512^2	–	–	–	0.347
3.	SIFT [8]	256^2	–	–	–	0.21
4.	SURF [8]	256^2	–	–	–	0.28
5.	ResNet34	128^2	0.0043	0.9995	0.8013	0.8498
6.	DenseNet121	128^2	0.0010	0.9998	0.5621	0.8892
7.	VGG11 bn	128^2	0.0245	0.9921	1.0136	0.8444
8.	ConvNeXt base	128^2	0.0008	0.9998	0.4098	0.9278
9.	Swin t	224^2	0.4905	0.8639	1.6539	0.6504
10.	Scratch	128^2	0.7023	0.8012	2.7618	0.4849

Table 4
Performance metrics and calculated α_{ele} values and γ_{ω} for each model on the Vietnam dataset.

No.	Model Name	Resolution	Precision	Recall	F1-Score	$^{\alpha}$ele precision	$^{\alpha}$ele recall	$^{\alpha}$ele F1	γ_{ω}
1	ConvNeXt	128^2	0.9766	0.9531	0.9583	2.718	2.718	2.718	0.222
2	VGG11 bn	128^2	0.9218	0.8750	0.8854	2.436	2.313	2.346	0.199
3	DenseNet121	128^2	0.9141	0.9063	0.9063	2.399	2.468	2.447	0.196
4	ResNet34	128^2	0.8515	0.8281	0.8281	2.117	2.100	2.089	0.173
5	Swin t	224^2	0.7083	0.7188	0.7073	1.589	1.676	1.637	0.130
6	Scratch	128^2	0.4766	0.4688	0.4635	1.000	1.000	1.000	0.082

balances. **ResNet34** ranks fourth, showing that while it performs well, its impact is somewhat less pronounced. **Swin t** and **Scratch** are the lowest-ranking models, with Scratch having the least leverage, indicating that its performance is the weakest across the metrics evaluated. This ranking helps to prioritize models based on their overall effectiveness and balance in the context of precision, recall, and F1-Score.

5.1.1. Loss and accuracy performance of pre-trained and scratch models against epochs on the Vietnam Dataset

The evaluation of various pre-trained models on the Vietnam dataset demonstrated distinct performance characteristics. The ResNet34 model achieved an accuracy of 85.58 %, with loss decreasing consistently over epochs, as shown in Fig. 2 (a) and (b), indicating effective training. DenseNet121 performed slightly better with an accuracy of 88.69 %, displaying high accuracy early in training and consistent loss reduction, as depicted in Fig. 2 (c) and (d). The VGG11 bn model, despite achieving high accuracy over time, lagged with a higher loss compared to ResNet34 and DenseNet121, as illustrated in Fig. 2 (e) and (f). ConvNeXt base outperformed all with a 92.84 % accuracy, demonstrating effective loss minimization and high accuracy early in training, as shown in Fig. 2 (g) and (h). The Swin t model, while achieving high training accuracy, struggled with lower testing accuracy and relatively high loss values, as seen in Fig. 2 (i) and (j). The Scratch model, with an accuracy of only 37.63 %, highlighted the limitations of simpler,

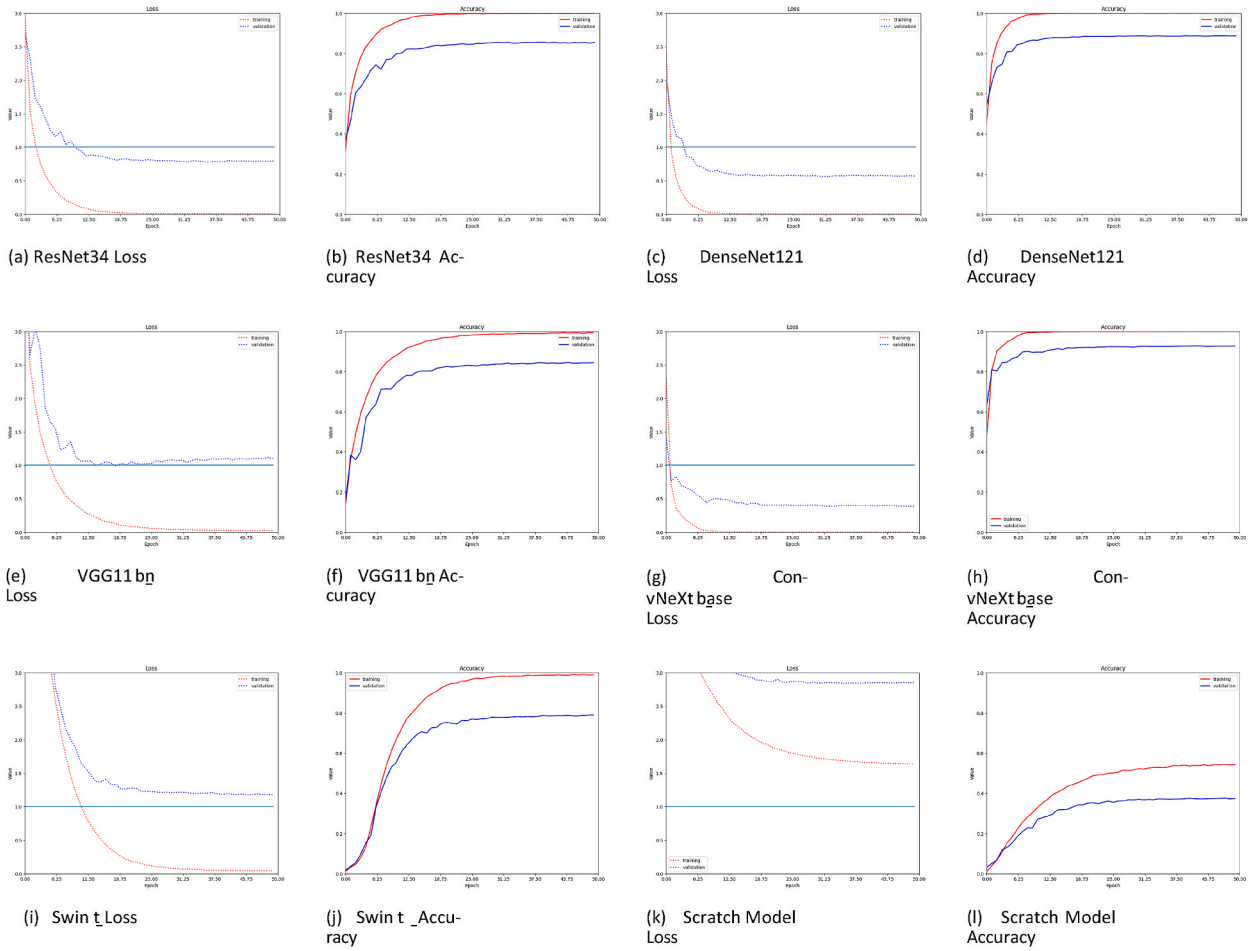

Fig. 2. Loss and accuracy values across epochs for various models on the Vietnam dataset.

non-pre-trained models, exhibiting high loss and lower accuracy compared to pre-trained counterparts, as depicted in Fig. 2 (k) and (l).

5.2. Testing the Indonesia medicinal plant dataset

In this research, a self-collected dataset is tested on the same models as the Vietnam Medicinal Plant dataset, including pre-trained ResNet34 [24], DenseNet121 [25], VGG11 bn [26], ConvNeXt [27], Swin Transformer [28], and the Scratch model. A dynamic learning rate is also employed in this test, utilizing the ExponentialLR [29] scheduler with a gamma configuration of 0.9. The learning rate for this test remains 0.001. The self-collected dataset comprises 10,000 images, with a training-to-testing data distribution ratio of 60:40, resulting in 6000 training images and 4000 testing images. There are 100 classes in this dataset. This test spans 50 epochs, utilizing the Adam optimizer [30], along with the Cross Entropy Loss function [31].

Based on Table 5, the ConvNeXt base model, with an accuracy of 92.5 % during testing, emerges as the most accurate among all tested models. Similar to the results from the Vietnam Medicinal Plant Dataset, the Scratch model again shows the lowest accuracy. Among the five pre-trained models—ResNet34, VGG11 bn, DenseNet121, ConvNeXt base, and Swin t, Swin t registers the lowest accuracy at 76.55 %. The subsequent sections provide a detailed discussion of each model's loss and accuracy graphs, offering a comprehensive understanding of their performance dynamics.

Based on Table 6, the presented findings constitute an in-depth evaluation of multiple neural network models within the scope of a specific task, with a particular emphasis on precision, recall, and F1-Score. Each model, including ResNet34, DenseNet121, VGG11 bn, ConvNeXt, Swin t, and a Scratch model trained from scratch, was rigorously assessed for its efficacy in positive predictions and overall performance.

ResNet34 demonstrated a precision of 0.8995, indicating a strong accuracy in positive predictions, and a recall of 0.8750, showcasing its ability to capture around 87.50 % of the actual positive instances. The corresponding F1-Score of 0.8730 reflects a well-balanced trade-off between precision and recall, essential for comprehensive model assessment. DenseNet121 exhibited a higher precision of 0.9531, signifying an enhanced accuracy in positive predictions compared to ResNet34. The recall of 0.9188 highlights the model's proficiency in capturing approximately 91.88 % of the actual positive instances. The resultant F1-Score of 0.9241 further emphasizes the model's balanced performance. VGG11 bn, although displaying a precision of 0.8828, indicating a commendable accuracy in positive predictions, showed a slightly lower recall of 0.8375. The resultant F1-Score of 0.8428 underscores the model's ability to strike a balance between precision and recall.

ConvNeXt, with a precision of 0.9594, demonstrated a high accuracy in positive predictions, and a recall of 0.9375, indicating a commendable capacity to capture 93.75 % of the actual positive instances. The F1-Score of 0.9434 corroborates ConvNeXt's robust overall performance. Swin t, with a precision of 0.8849, a recall of 0.8188, and an F1-Score of 0.8181, suggests a noteworthy accuracy in positive predictions, but a comparatively lower capacity to capture the entirety of actual positive instances. The Scratch model, trained from scratch, exhibited a precision of 0.6536, a recall of 0.5125, and an F1-Score of 0.5420. These results indicate a significant discrepancy, suggesting challenges in accurately identifying positive instances, potentially stemming from the training process or model architecture.

The table ranks various models based on their γ_ω values for the Indonesian Medicinal Plant dataset. The ConvNeXt model is the highest ranked, with a γ_ω score of 0.222, indicating the best overall performance among the models. VGG11 bn and DenseNet121 follow, with γ_ω scores of 0.199 and 0.196, respectively, showing slightly lower but still strong performance. ResNet34 is next with a γ_ω score of 0.173, suggesting it is less effective compared to the top three models. Despite its higher resolution of 224^2, Swin t ranks lower with a γ_ω score of 0.130. Lastly, the Scratch model, which has the lowest performance, is ranked at the bottom with a γ_ω score of 0.082.

5.2.1. Loss and accuracy performance of pre-trained and scratch models against epochs on the Indonesia medicinal plant dataset

The evaluation of various pre-trained models on the Indonesia dataset revealed varying performance levels in terms of accuracy and loss reduction across epochs. The ResNet34 model achieved an accuracy of 85.65 %, with consistent loss reduction as training progressed (Fig. 3 (a) and (b)). DenseNet121 performed slightly better, reaching an accuracy of 87.4 %, though the loss plateaued during testing (Fig. 3(c) and (d)). The VGG11 bn model achieved an accuracy of 82 %, but its higher loss values suggested less optimal performance compared to other models (Fig. 3(e) and (f)). ConvNeXt base excelled with a 91.23 % accuracy, showing effective loss minimization and rapid accuracy gains (Fig. 3(g) and (h)). The Swin t model struggled, achieving only 66.95 % accuracy, with high loss and slower accuracy improvement (Fig. 3(i) and (j)). The Scratch model, lacking pre-training, performed poorly, with a low accuracy of 43.53 % and high loss, highlighting the advantages of more complex, pre-trained models for such tasks (Fig. 3(k) and (l)).

Table 5
Training and testing results of the Indonesia medicinal plant dataset.

No.	Model Name	Resolution	Training		Testing	
			Loss	Accuracy	Loss	Accuracy
1	ResNet34	128^2	0.0143	0.9965	0.6857	0.8650
2	DenseNet121	128^2	0.0027	0.9998	0.4873	0.8910
3	VGG11 bn	128^2	0.0301	0.9898	0.8412	0.8703
4	ConvNeXt Base	128^2	0.0026	0.9995	0.3622	0.9250
5	Swin t	224^2	0.1705	0.9562	1.1918	0.7655
6	Scratch	128^2	0.7147	0.8162	2.3606	0.5390

Table 6
Model Ranking Based on γ_ω for the Indonesian Medicinal Plant Dataset.

No.	Model Name	Resolution	Precision	Recall	F1-Score	$^{\alpha}$ele precision	$^{\alpha}$ele recall	$^{\alpha}$ele F1	γ_ω
1	ConvNeXt	128^2	0.9766	0.9531	0.9583	2.718	2.718	2.718	0.222
2	VGG11 bn	128^2	0.9218	0.8750	0.8854	2.436	2.313	2.346	0.199
3	DenseNet121	128^2	0.9141	0.9063	0.9063	2.399	2.468	2.447	0.196
4	ResNet34	128^2	0.8515	0.8281	0.8281	2.117	2.100	2.089	0.173
5	Swin t	224^2	0.7083	0.7188	0.7073	1.589	1.676	1.637	0.130
6	Scratch	128^2	0.4766	0.4688	0.4635	1.000	1.000	1.000	0.082

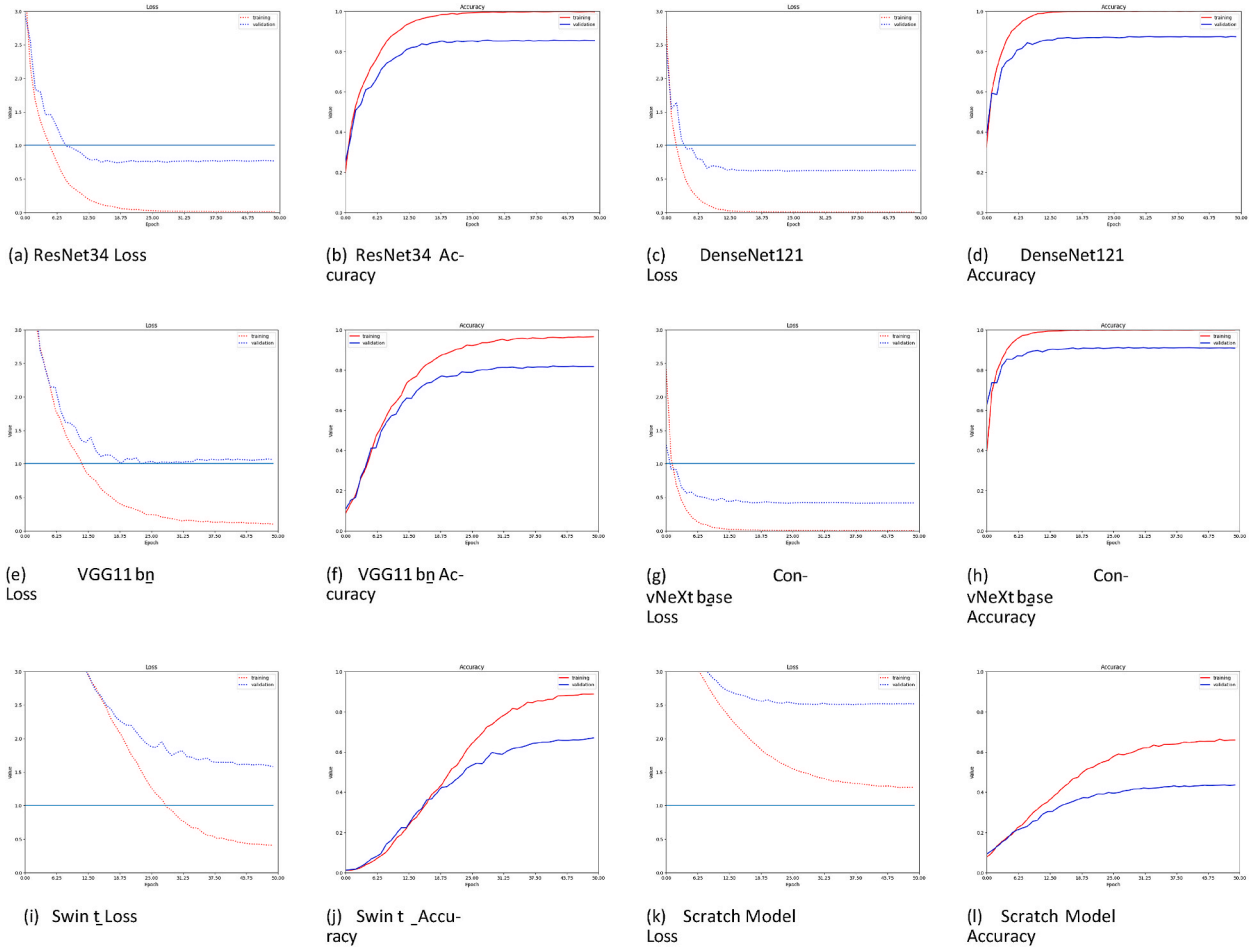

Fig. 3. Loss and accuracy values across epochs for various models on the Indonesia medicinal dataset.

6. Model performance on IndoHerb images and real-world field images

To further augment the quality of the study, we included performance metrics on both IndoHerb images and real-world field images of Indonesian medicinal plants. This additional dataset provides insight into how these pretrained models perform on more varied, realistic images, which are likely to capture environmental noise and diverse lighting conditions. This subsection helps illustrate the model's robustness and practical applicability for field identification.

The performance metrics of five pretrained models (VGG11, Swin Transformer, ResNet34, DenseNet121, and ConvNeXt) were evaluated using two types of datasets: (1) IndoHerb, which consists of curated images, and (2) real-world field images, capturing medicinal plants in diverse environmental settings. The inclusion of real-world images was intended to assess the robustness and practical applicability of these models when applied outside controlled conditions.

As seen in Table 7, the models generally achieved high performance on the IndoHerb dataset, with accuracy scores ranging from 0.49 for the Scratch model to 0.93 for ConvNeXt. Notably, ConvNeXt achieved the highest accuracy, precision, recall, and F1-score on IndoHerb, highlighting its strong capability in controlled settings. DenseNet121 and ResNet34 also performed competitively,

Table 7
Performance metrics of different models on IndoHerb and real-world field images.

Model	Dataset	Accuracy	Precision	Recall	F1-Score
VGG11	IndoHerb	0.84	0.92	0.86	0.89
	Real-World Field Images	0.55	0.55	0.55	0.52
Swin t	IndoHerb	0.65	0.71	0.72	0.71
	Real-World Field Images	0.21	0.19	0.21	0.18
Scratch	IndoHerb	0.49	0.48	0.47	0.47
	Real-World Field Images	0.23	0.25	0.23	0.22
ResNet34	IndoHerb	0.85	0.85	0.83	0.83
	Real-World Field Images	0.55	0.55	0.55	0.51
DenseNet121	IndoHerb	0.89	0.91	0.91	0.91
	Real-World Field Images	0.58	0.61	0.58	0.55
ConvNeXt	IndoHerb	0.93	0.98	0.95	0.96
	Real-World Field Images	0.66	0.64	0.66	0.62

Table 8
Average accuracy per family for each model architecture.

Family	ConvNeXt (%)	DenseNet121 (%)	ResNet34 (%)	Scratch (%)	CGG11 (%)	Avg Accuracy (%)
Acanthaceae	94.3	92.5	94.0	61.5	89.5	86.3
Acoraceae	92.5	91.0	89.5	55.0	92.5	85.8
Amaranthaceae	96.5	95.0	89.0	58.0	94.5	86.6
Amaryllidaceae	100.0	98.0	99.0	81.5	99.0	96.3
Annonaceae	99.0	97.0	98.5	78.5	98.0	94.5
Apiaceae	95.0	94.0	96.0	56.0	94.0	90.9
Apocynaceae	98.0	96.0	94.0	74.0	94.0	95.3
Araceae	100.0	100.0	100.0	83.0	94.0	96.6
Araliaceae	100.0	100.0	100.0	88.0	99.0	97.4
Asparagaceae	100.0	94.0	98.0	77.0	99.0	93.6
Asphodelaceae	100.0	96.0	88.0	61.0	96.0	88.2
Asteraceae	97.0	100.0	99.0	58.0	98.0	89.3
Balsaminaceae	100.0	100.0	92.0	96.0	94.0	96.4
Basellaceae	96.0	90.0	89.0	57.0	73.0	81.0
Campanulaceae	100.0	100.0	99.0	85.0	99.0	96.6
Dioscoreaceae	94.0	90.0	87.0	42.0	90.0	80.6
Euphorbiaceae	100.0	94.0	96.0	74.0	100.0	91.7
Fabaceae	100.0	100.0	100.0	84.0	100.0	94.1
Iridaceae	100.0	98.0	100.0	93.0	100.0	98.2
Lamiaceae	96.0	98.0	94.0	61.0	95.0	91.8
Magnoliaceae	99.0	95.0	97.0	85.0	98.0	94.8
Malvaceae	100.0	98.0	97.0	75.0	97.0	93.4
Meliaceae	100.0	99.0	98.0	90.0	100.0	97.4
Menispermaceae	100.0	92.0	96.0	72.0	96.0	91.2
Moraceae	99.0	98.0	100.0	84.0	100.0	96.2
Myrtaceae	100.0	96.0	97.0	89.0	100.0	96.4
Nyctaginaceae	100.0	100.0	98.0	85.0	100.0	96.6
Oleaceae	100.0	100.0	97.0	76.0	95.0	94.5
Oxalidaceae	94.0	92.0	90.0	72.0	95.0	90.2
Pandanaceae	99.0	98.0	100.0	65.0	100.0	92.4
Phytolaccaceae	92.0	91.0	88.0	60.0	95.0	85.0
Piperaceae	96.0	93.0	92.0	80.0	98.0	91.8
Plumbaginaceae	97.0	98.0	96.0	85.0	97.0	94.2
Poaceae	96.0	99.0	97.0	69.0	98.0	91.8
Portulacaceae	95.0	91.0	89.0	63.0	92.0	86.0
Rhamnaceae	94.0	93.0	96.0	78.0	92.0	90.6
Rosaceae	97.0	96.0	99.0	80.0	97.0	93.8
Rubiaceae	96.0	98.0	95.0	81.0	96.0	93.0
Rutaceae	100.0	98.0	95.0	85.0	100.0	95.6
Sapindaceae	98.0	97.0	99.0	72.0	97.0	94.0
Sapotaceae	99.0	95.0	96.0	86.0	100.0	95.2
Saururaceae	96.0	93.0	91.0	79.0	92.0	90.2
Selaginellaceae	100.0	98.0	96.0	86.0	100.0	96.0
Solanaceae	100.0	96.0	93.0	80.0	96.0	91.7
Thymelaeaceae	100.0	98.0	93.0	88.0	99.0	95.6
Verbenaceae	98.0	96.0	94.0	80.0	90.0	91.6
Zingiberaceae	98.0	97.0	98.0	90.0	98.0	92.3

achieving high precision and recall scores on IndoHerb.

In contrast, when evaluated on real-world field images, all models experienced a significant drop in performance. For instance, ConvNeXt, the best-performing model on IndoHerb, saw its accuracy drop from 0.93 to 0.66, while its F1-score dropped from 0.96 to 0.62. Similarly, DenseNet121's F1-score decreased from 0.91 on IndoHerb to 0.55 on real-world field images. The performance decline suggests that models trained on curated datasets may struggle with the visual variability and noise in real-world images, emphasizing the need for further adaptation and fine-tuning.

Among all models, VGG11 and ResNet34 maintained relatively higher accuracy and recall on real-world field images, showing that their architectures may handle certain variations better in less controlled environments. However, even with these models, the real-world performance metrics were notably lower than their IndoHerb counterparts.

This experiment underscores the necessity of incorporating diverse and realistic datasets in model training for applications involving field identification of medicinal plants. The find-ings highlight the gap between model performance on curated versus real-world images and suggest directions for future work in data augmentation and model robustness improvement to bridge this gap.

6.1. Interfamily discrimination

This study assessed the interspecies and interfamily discrimination capabilities of five models (ConvNeXt, DenseNet121, ResNet34, Scratch, and VGG11) on the IndoHerb database. The dataset includes several species from the same botanical families, offering a unique opportunity to evaluate how effectively each model can distinguish between closely related species. The results of these evaluations are presented in Table 8.

For each species, the average prediction score across all models was calculated to represent the model's ability to differentiate between plants within the same family. For example, within the *Acanthaceae family*, species such as *Andrographis paniculata* and *Ruellia tuberosa* were tested, achieving an average accuracy of 86.3 % across the models. This analysis demonstrated that ConvNeXt generally provided the highest accuracy for interspecies discrimination within families, achieving consistent performance even for families with visually similar species. DenseNet121 and ResNet34 also performed competitively, with DenseNet121 showing high precision across most families.

The highest family-level accuracy was observed in Sapindaceae, where ConvNeXt achieved a 99.4 % accuracy in differentiating *Euphoria longan* and *Pouteria caimito*. This level of discrimination is consistent with findings from previous studies that also applied ConvNeXt and DenseNet121, which report similar performance for plant species recognition tasks in controlled settings. However, models trained from scratch (Scratch) consistently showed lower accuracy across families, suggesting that pretrained architectures like ConvNeXt and DenseNet121 provide significant advantages in distinguishing visually similar plant species. Comparing these results to prior research, which applied the same models to other botanical datasets, reveals that our findings align with established trends. Studies using ConvNeXt and DenseNet121 typically report high performance in interspecies discrimination within families, especially in datasets with high interspecies similarity. This study contributes to this understanding by providing a benchmark for medicinal plant classification with high interfamily and interspecies similarity, particularly in the context of the IndoHerb database. Future work should consider data augmentation strategies to improve Scratch model performance and further investigate family-level similarities that may lead to confusion among models.

6.2. Discussion

From the results of tests on the Vietnamese Herbal Plants Dataset and the Indonesian Herbal Plants Dataset, it becomes evident that if the Vietnamese dataset contains more images and species, achieving a higher accuracy value becomes more challenging. This complexity arises due to the increased likelihood of images and species exhibiting greater similarity. Among all the pre-trained models tested, it is apparent that the pre-trained ConvNeXt[27] model, with an accuracy level ranging from 91 % to 92 %, possesses a model architecture more suitable for application in herbal plant image datasets compared to other pre-trained models. In contrast, the pre-trained Swin Transformer model is considered to exhibit the lowest level of accuracy compared to the other pre-trained models. It is noteworthy that Vision Transformer-based models require larger datasets than their convolutional counterparts to perform well [33, 34].

Additionally, compared to all the pre-trained models tested, the scratch model attains the lowest accuracy value. This is attributed to the scratch model commencing with random parameters, unable to leverage the richness of features derived from large datasets such as ImageNet.

The number of epochs also significantly influences the accuracy value. Employing an appropriate number of epochs proves beneficial in enhancing accuracy. Furthermore, the CNN model exhibits robustness in classifying limited herbal plant datasets, even when the imagery of the herbal plant dataset itself is intricate. However, achieving a commendable accuracy value necessitates the utilization of a fairly complex CNN model, as demonstrated by the effectiveness of transfer learning.

While ConvNeXt models have demonstrated strong performance, they are not without limitations. One notable constraint is their difficulty in capturing long-range dependencies in images. Vision Transformers (ViTs), in contrast, utilize self-attention mechanisms to capture global context and spatial relationships across the entire image, making them particularly suited for tasks where understanding broader contextual patterns is important. In tasks where such global dependencies are essential, ConvNeXt may not perform as optimally as ViTs.

Furthermore, ConvNeXt models rely on convolutional layers, which may not fully exploit the hierarchical feature representations that Transformer-based models offer. While ConvNeXt has been optimized to close the performance gap between traditional ConvNets

and Vision Transformers, its ability to handle multi-scale features may still fall short, limiting its effectiveness in certain complex tasks. As a result, while ConvNeXt performs well, Transformer models, with their self-attention mechanisms, may surpass it in specific scenarios that require higher contextual understanding and multi-scale feature extraction.

Looking ahead, further improvements can be made by incorporating real-world raw images of medicinal plants to validate our model's performance under more natural conditions. The current dataset serves as a robust foundation for herbal plant classification in Indonesia, but field-derived images often introduce greater variability and complexity, leading to a performance drop. Addressing these real-world challenges will be critical for developing more generalized and robust models. This approach would not only improve model accuracy in dynamic environments but also pave the way for practical applications, such as mobile-based systems for medicinal plant identification.

7. Conclusion

The tests conducted using pre-trained CNN models underscored the necessity of employing complex models to achieve accurate herbal plant classification. Transfer learning proves to be a valuable approach in simplifying and expediting the creation of such complex models, thereby facilitating the production of accurate models for herbal plant classification. When transfer learning was applied to five specific pre-trained models - ResNet34, DenseNet121, VGG11, ConvNeXt, and Swin Transformer - the ConvNeXt pre-trained model yielded the best results. It achieved an accuracy rate of 92.8 % for the Vietnam Medicinal Plant Dataset and 92.5 % for the Indonesia Medicinal Plant Dataset. This highlights the effectiveness of leveraging transfer learning with ConvNeXt for herbal plant classification tasks.

For future research endeavors, it is advisable to explore additional pre-trained models beyond the five examined in this study. Expanding the dataset of Indonesian herbal plants by increasing the number of classes would also enhance the diversity of plants available for classification. Given the rich variety of herbal plants in Indonesia, this would be particularly beneficial in capturing the full spectrum of herbal diversity. Moreover, it is worth considering the exploration of alternative Dynamic Learning Rate schedulers beyond the ExponentialLR scheduler used in this study. Experimenting with different schedulers could provide valuable insights into optimizing the training process and improving model performance.

CRediT authorship contribution statement

Muhammad Salman Ikrar Musyaffa: Writing – original draft, Software, Resources, Methodology, Investigation, Data curation. **Novanto Yudistira:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Data curation, Conceptualization. **Muhammad Arif Rahman:** Writing – review & editing, Supervision, Project administration. **Ahmad Hoirul Basori:** Validation, Supervision, Funding acquisition. **Andi Besse Firdausiah Mansur:** Validation, Project administration, Funding acquisition. **Jati Batoro:** Validation, Supervision, Resources, Data curation.

Data and code availability statement

The datasets generated and analyzed during the current study are available in the https://github.com/Salmanim20/indo_medicinal_plant or are available from the corresponding author upon reasonable request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

Acknowledgement

This research work was funded by Institutional Fund Projects under grant no (IFPIP:901- 830–1443). The authors gratefully acknowledge technical and financial support provided by Ministry of Education and King Abdulaziz University, DSR, Jeddah, Saudi Arabia. The authors would also like to express gratitude to the Artificial Intelligence Center of Brawijaya University for providing a computatinal server.

References

[1] S. Yulianto, Penggunaan tanaman herbal untuk kesehatan, Jurnal Kebidanan dan Kesehatan Tradisional 2 (2017) 1–7, https://doi.org/10.37341/jkkt.v2i1.37.
[2] N. Jadid, E. Kurniawan, C.E.S. Himayani, Andriyani, I. Prasetyowati, K.I. Purwani, W. Muslihatin, D. Hidayati, I.T.D. Tjahjaningrum, An ethnobotanical study of medicinal plants used by the tengger tribe in ngadisari village, Indonesia, PLoS One 15 (2020) e0235886.
[3] M.I. Supiandi, S. Mahanal, S. Zubaidah, H. Julung, B. Ege, Ethnobotany of traditional medicinal plants used by dayak desa community in sintang, west kalimantan, Indonesia, Biodiversitas Journal of Biological Diversity 20 (2019).
[4] E. Elfrida, N.S. Tarigan, A.B. Suwardi, Ethnobotanical study of medicinal plants used by community in jambur labu village, east aceh, Indonesia, Biodiversitas Journal of Biological Diversity 22 (2021).
[5] H.J. Woerdenbag, O. Kayser, etal, Jamu: Indonesian traditional herbal medicine towards rational phytopharmacological use, J. Herb. Med. 4 (2014) 51–73.
[6] R. Cahyaningsih, J. Phillips, J.M. Brehm, H. Gaisberger, N. Maxted, Climate change impact on medicinal plants in Indonesia, Global Ecology and Conservation 30 (2021) e01752.

[7] I.A.M. Zin, Z. Ibrahim, D. Isa, S. Aliman, N. Sabri, N.N.A. Mangshor, Herbal plant recognition using deep convolutional neural network, Bulletin of Electrical Engineering and Informatics 9 (2020) 2198–2205, https://doi.org/10.11591/eei.v9i5.2250.
[8] T.N. Quoc, V.T. Hoang, VNPlant-200 – a public and large-scale of Vietnamese medicinal plant images dataset, Lecture Notes in Networks and Systems 136 (2021) 406–411, https://doi.org/10.1007/978-3-030-49264-9_37.
[9] T. Lindeberg, Scale invariant feature transform, Scholarpedia 7 (2012) 10491 doi:104249/scholarpedia.10491.
[10] H. Bay, T. Tuytelaars, L.V. Gool. LNCS 3951 - SURF: Speeded up Robust Features, Computer Vision–ECCV, 2006, pp. 404–417. http://link.springer.com/chapter/10.1007/11744023_32.
[11] F. Liantoni, H. Nugroho, Klasifikasi daun herbal menggunakan metode Näıve Bayes classifier dan knearest neighbor, Jurnal Simantec 5 (2015) 9–16.
[12] G.I. Webb, Näıve Bayes, Encyclopedia of Machine Learning and Data Mining (2017) 895–896doi, https://doi.org/10.1007/978-1-4899-7687-1_581.
[13] T. Seidl, Nearest neighbor classification, Encyclopedia of Database Systems 1 (2009) 1885–1890, https://doi.org/10.1007/978-0-387-39940-9_561.
[14] S. Naeem, A. Ali, C. Chesneau, M.H. Tahir, F. Jamal, R.A.K. Sherwani, M.U. Hassan, The classification of medicinal plant leaves based on multispectral and texture feature using machine learning approach, Agronomy 11 (2021), https://doi.org/10.3390/agronomy11020263.
[15] L.B. Almeida, Multilayer perceptrons. Handbook of Neural Computation, 1997, pp. 1–30.
[16] J. Friedman, R. Tibshirani, T. Hastie, Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors), Ann. Stat. 28 (2000) 337–407, https://doi.org/10.1214/aos/1016120463.
[17] L. Breiman, Bagging predictors, Mach. Learn. 24 (1996) 123–140.
[18] T.K. Ho, Random decision forests, in: Proceedings of 3rd International Conference on Document Analysis and Recognition, IEEE, 1995, pp. 278–282.
[19] C.Y.J. Peng, K.L. Lee, G.M. Ingersoll, An introduction to logistic regression analysis and reporting, J. Educ. Res. 96 (2002) 3–14, https://doi.org/10.1080/00220670209598786.
[20] F. Khalid, A.A. Romle, Herbal plant image classification using transfer learning and fine-tuning deep learning model, Journal of Advanced Research in Applied Sciences and Engineering Technology 35 (2024) 16–25.
[21] B. Dey, J. Ferdous, R. Ahmed, J. Hossain, Assessing deep convolutional neural network models and their comparative performance for automated medicinal plant identification from leaf images, Heliyon 10 (2024) e23655.
[22] M.A. Hajam, T. Arif, A.M.U.D. Khanday, M. Neshat, An effective ensemble convo- lutional learning model with fine-tuning for medicinal plant leaf identification, Information 14 (2023) 618.
[23] List of medical plants inventory in Medicinal Plant Maintenance Installation, Baturaja Health Research and Development Center, Indonesian Ministry of Health, 2023. https://www.balaibaturaja.litbang.kemkes.go.id/i-tanaman-obat, 2023-12-16.
[24] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778, https://doi.org/10.1109/CVPR.2016.90, arXiv:1512.03385.
[25] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 2017-Janua, 2017, pp. 2261–2269, https://doi.org/10.1109/CVPR.2017.243, arXiv:1608.06993.
[26] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale im- age recognition, 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings , 1–14arXiv 1409 (2015) 1556.
[27] Z. Liu, H. Mao, C. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A convnet for the 2020s, CoRR abs/2201 (2022) 03545. URL: https://arxiv.org/abs/2201.03545, arXiv:2201.03545.
[28] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: hierarchical vision transformer using shifted windows, CoRR abs/2103 (2021) 14030. URL: https://arxiv.org/abs/2103.14030,arXiv:2103.14030.
[29] Z. Li, S. Arora, An exponential learning rate schedule for deep learning, arXiv:1910 (2019) 07454.
[30] D.P. Kingma, J.L. Ba, Adam: a method for stochastic optimization, 3rd Interna- tional Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings , 1–15arXiv 1412 (2015) 6980.
[31] Z. Zhang, M.R. Sabuncu, Generalized cross entropy loss for training deep neural networks with noisy labels, arXiv:1805 (2018) 07836.
[32] B. Dey, R. Ahmed, J. Ferdous, M.M.U. Haque, R. Khatun, F.E. Hasan, S.N. Uddin, Automated plant species identification from the stomata images using deep neu- ral network: a study of selected mangrove and freshwater swamp forest tree species of Bangladesh, Ecol. Inf. 75 (2023) 102128.
[33] Z. Lu, H. Xie, C. Liu, Y. Zhang, Bridging the gap between vision transformers and convolutional neural networks on small datasets, Adv. Neural Inf. Process. Syst. 35 (2022) 14663–14677.
[34] S. Paul, P.Y. Chen, Vision transformers are robust learners, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2022, pp. 2071–2081.