Baghdad Science Journal

Volume 22 | Issue 3 Article 30

3-28-2025

Medicinal Plant Leaf Classification using Deep Learning and Vision Transformers

Shahriar Hossain
Department of Computer Science, George Mason University, Fairtax, USA, shossa4@gmu.edu

Rizbanul Hasan
Department of Computer Science, George Mason University, Fairtax, USA, rhasan7@gmu.edu

Jia Uddin
Department of Artificial Intelligence and Big Data, Endicott College, Woosong University, Daejeon, South Korea, jia.uddin@wsu.ac.kr

How to Cite this Article

Hossain, Shahriar; Hasan, Rizbanul; and Uddin, Jia (2025) "Medicinal Plant Leaf Classification using Deep Learning and Vision Transformers," *Baghdad Science Journal*: Vol. 22: Iss. 3, Article 30.
DOI: https://doi.org/10.21123/bsj.2024.10844

BAGHDAD SCIENCE JOURNAL 2025;22(3):1065–1076

RESEARCH ARTICLE

Medicinal Plant Leaf Classification using Deep Learning and Vision Transformers

Shahriar Hossain [1], **Rizbanul Hasan** [1], **Jia Uddin** [2,*]

[1] Department of Computer Science, George Mason University, Fairtax, USA
[2] Department of Artificial Intelligence and Big Data, Endicott College, Woosong University, Daejeon, South Korea

ABSTRACT

Identification of medicinal plant leaves is very crucial as their cultivation and production are essential for the medicine industry. Many different classes of medicinal leaves look identical but serve different purposes in the medicine industry and have different remedies for different diseases. Hence it is imperative to use methods that are automated, faster, and produce good accuracy. Cutting-edge models have been trained to discern the subtle distinctions between various species of leaves, accounting for a myriad of factors such as leaf texture, shape, and color variations, which are often imperceptible to the human eye. In this research, Transfer learning (TL) based VGG16 and Vision Transformer (ViT) models such as ConvMixer and Compact Convolutional Transformer (CCT) are implemented for the classification of medicinal leaf images using a dataset of 38066 leaf images having 10 different classes. The proposed customized Convolutional Neural Network (CNN) and hybrid CNN-ViT models both have a very low number of parameters compared to the other models in comparison making them light and capable of being less computationally expensive. In the experimental evaluation, all the results are collected for 30 epochs. VGG16, CCT, and ConvMixer produce AUC scores of 0.50, 0.79, and 0.50, respectively for the dataset while the proposed CNN and hybrid model gave AUC scores of 0.83 and 0.74, respectively. In addition, a hybrid denoising approach with Wavelet thresholding and Gaussian blurring is utilized to minimize the noises in the images by retaining the original image quality.

Keywords: Classification, CNN-ViT, Medicinal plant leaf, Transfer learning, Vision transformer

Introduction

Medicinal plants play a vital role in healthcare. These plants offer a rich source of natural compounds with healing properties with relatively fewer side effects. As well as being a source of natural supplements these plants are the foundation of many pharmaceutical drugs. As continue to expand our knowledge in modern medicine, the knowledge about the medicinal plants received from predecessors should be continued as they are integral to advancing healthcare practices globally. For this, it is extremely important to develop modern, fast, and accurate techniques [1] to identify different types of medicinal plants.

In this regard, Deep learning (DL) and ViTs have revolutionized the idea of image classification tasks by introducing powerful tools for representation learning and feature extraction. DL architectures, particularly CNNs, have been crucial for image classification advancements. [2–4] These models can learn hierarchical features from raw pixel data, enabling them to understand complex patterns and representations. With the advancement of image classification, ViTs have emerged as an alternative architecture. ViTs utilize a self-attention mechanism to capture long-range dependencies and relationships among different parts of the image. [5] This attention-based approach allows ViTs to model global context and

Received 4 February 2024; revised 24 May 2024; accepted 26 May 2024.
Available online 28 March 2025

* Corresponding author.
E-mail addresses: shossa4@gmu.edu (S. Hossain), rhasan7@gmu.edu (R. Hasan), jia.uddin@wsu.ac.kr (J. Uddin).

https://doi.org/10.21123/bsj.2024.10844

have improved performance in image classification tasks.[5]

In the case of medicinal plants, specialists still use the manual method (visual identification) for identifying or classification of medicinal plants. However, it is very difficult to identify various types of medicinal plants due to similarities with some other plants. It is also a very time-consuming and tedious process. So, it is important to use methods that are automated, faster, and can produce good accuracy. As the characteristics of a plant's leaf can be easily extracted and evaluated, therefore, it is commonly used as the primary method for identifying medicinal plants. So, to address this issue, DL and ViTs are used to classify and identify different medicinal plants easily and with high accuracy. The main contributions of this paper can be summarized as follows.

- Introduced a hybrid denoising algorithm using Wavelet thresholding and Gaussian blurring that is capable of both feature preservation and noise reduction.
- Three state-of-the-art (SOTA) TL and ViT models in the form of VGG16, CCT, and ConvMixer were implemented having their parameters tweaked to increase their efficacy.
- Through rigorous experiments two unique models were presented in the form of a lightweight CNN model and a hybrid CNN-ViT model having a very low number of parameters.
- Additionally, compared the effectiveness of the proposed models with the SOTA models across various results metrics such as accuracy curves, confusion matrices, classification reports, and ROC curves and found the models to produce comparable results with a very low number of parameters against the SOTA models having a very large number of parameters.

Rest of the paper is organized as follows. Section 2 presents the background study and proposed methodology in Section 3. Experimental results analysis is in Section 4. Finally, conclude the paper in Section 5.

Literature review

This segment discusses some of the relevant research on the classification of medicinal images. The research paper[3] presents DOLG-NeXt, an advanced encoder-decoder architecture designed for the segmentation of biomedical images. It improves feature extraction by incorporating SE-Net-driven ConvNeXt in the encoder and uses a fusion module in the decoder to capture detailed contextual features. Additionally, it employs a lightweight attention network to improve the decoding process. DOLG-NeXt demonstrates superior performance compared to contemporary models on a variety of biomedical image datasets, achieving dice coefficient scores, such as 95.10% in CVC-ClinicDB, 95.80% in ISBI 2012, 94.77% in 2018 Data Science Bowl, and 84.88% in the DRIVE dataset. This paper[6] introduces a Deep CNN called DCPLD-CNN, leveraging the power of DL models to automatically extract features from images. The research explores eight pre-trained CNN architectures, including DenseNet121, NasNet-Large, VGG16, VGG19, ResNet50, InceptionV3, Inception-ResNetV2, and Xception, fine-tuning them through TL to identify diseases in cotton plants and leaves. Extra dense layers are added to the last layers of these models. Image Data Augmentation (IDA) techniques are utilized to expand the training data, enhancing model generalization and reducing overfitting. The proposed DCPLD-CNN model achieves an impressive accuracy of 98.77% in disease recognition in cotton plants and leaves, with a customized DenseNet121 model achieving the highest accuracy of 98.60% among all pre-trained architectures.

The article[7] reviews the performance and predictability of various machine learning (ML) and DL algorithms employed in recent times for plant categorization using leaf images. It discusses image processing techniques applied to certain classifiers to identify leaves and extract crucial leaf features. According to the review CNN-based approaches generally outperform alternative classifiers, while ANN classifiers are gaining popularity due to their high accuracy rates. This study[8] is based on classifying medicinal plant leaves using an ML approach. The authors used six varieties of medicinal plant leaves-based datasets where each plant had 600 samples of digital image and 600 samples of multispectral dataset. Feature extraction encompassed 65 fused features, combining texture, spectral, and gray-level run-length matrix features. Five ML classifiers were evaluated, with the multilayer perceptron (MLP) achieving the highest accuracy of 99.01%.

The aim of the proposed work[9] is to provide a way to detect the diseases of tomato leaves with images using the DL pre-trained model ResNet. The dataset contains 6,888 images about six tomato leaf diseases and healthy leaf images. They used a CNN model named ResNet101 for feature learning and image classification. Their proposed model gave a 98.8% accuracy. Hassoon et al.[1] utilized Feed Foreword Neural Network for Potato diseases classification. This research[10] demonstrates the robustness of the deep neural networks model for automatically identifying different kinds of medicinal leaves using several DL techniques. The authors developed a

Table 1. Overview of the reviewed sources.

Task	Datasets	Classifiers	Accuracy
Segmentation of biomedical images[11]	CVC-ClinicDB; ISBI 2012; Data Science Bowl 2018	DOLG-NeXt	95.10%, 95.80%, 94.77%, 84.88%
Disease identification in cotton plants[6]	*Not specified*	DCPLD-CNN; DenseNet121; NasNet-Large; VGG16; VGG19; ResNet50; InceptionV3; Inception-ResNetV2; Xception	98.77%
Plant categorization using leaf images[7]	*Not specified*	Various ML and DL algorithms	*Not specified*
Classifying medicinal plant leaves[8]	Six medicinal plant leaves datasets	MLP	99.01%
Detecting tomato leaf diseases[9]	Dataset consisting of 6888 images	ResNet101	98.8%
Potato diseases classification[1]	*Not specified*	Feed forward neural network	*Not specified*
Identification of medicinal leaves[10]	32,312 images of 30 different leaves	CNN; VGG16; MobileNet; Xception; InceptionResNetV2	99.8%
Indoor and outdoor image classification[3]	*Not specified*	GoogleNet; MobileNetv2	*Not specified*

mobile application with CNN to identify 30 types of medicinal leaves. A total of 32, 312 images of 30 different leaves were used in this study. TL models like VGG16, MobileNet, Xception, and InceptionResNetV2 were also implemented. They got the highest accuracy of 99.8% using InceptionResNetV2 with a smaller number of epochs (4). In,[3] Jassim et al. used pre-trained GoogleNet and MobileNetv2 for indoor and outdoor image classifications. A brief summary of the state-of-the-art models are presented in Table 1.

Methodology

This section presents the details methodology of proposed model.

Dataset

Image augmentation: To validate the model, the dataset[12] is used, which consists of 2,029 original medicinal leaf images, along with an additional 38,606 augmented images. The authors used the Keras ImageDataGenerator class to increase the image count. Various image enhancements, such as a 60-degree rotation, 10% zoom, flipping both horizontally and vertically, adjusting brightness (between 0.2 and 0.3), 15% shearing, and shifting height and width by 10% were applied. All enhanced images were resized to a resolution of 512 × 512 pixels.

Image partition: This dataset has a total of 10 classes. The augmented dataset is divided into training, testing, and validation with 80:10:10 ratios for each directory. A sample of data from each of the classes is shown in Fig. 1.

Data preprocessing

Conventional techniques frequently concentrate on the spatial or frequency domain; nonetheless, hybrid strategies[13] have demonstrated the potential to strike a balance between feature preservation and noise reduction. The OpenCV and PyWavelets libraries are used in the Python implementation of the hybrid denoising algorithm. Wavelet thresholding[14] and Gaussian blurring[15] are the two primary steps in the procedure. A sample of data before and after preprocessing is shown in Fig. 2.

Gaussian blurring: The R, G, and B color channels of an image are each subjected to a Gaussian blur to smooth out fluctuations and lower high-frequency noise. The Gaussian filter G (x, y) is defined by the following Eq. (1).

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (1)$$

In the Eq. (1), x and y are the distances from the origin in the horizontal and vertical axes, respectively and σ is the standard deviation of the Gaussian distribution.

Wavelet thresholding: Wavelet Decomposition and thresholding: In the next step, the Gaussian blurred image is broken down into approximation coefficients (cA) and detail coefficients (cH, cV, cD), which stand for horizontal, vertical, and diagonal features, respectively, by the discrete wavelet transform. These coefficients are then thresholded to reduce noise.

Wavelet reconstruction: To obtain the denoised image, wavelet reconstruction is carried out following

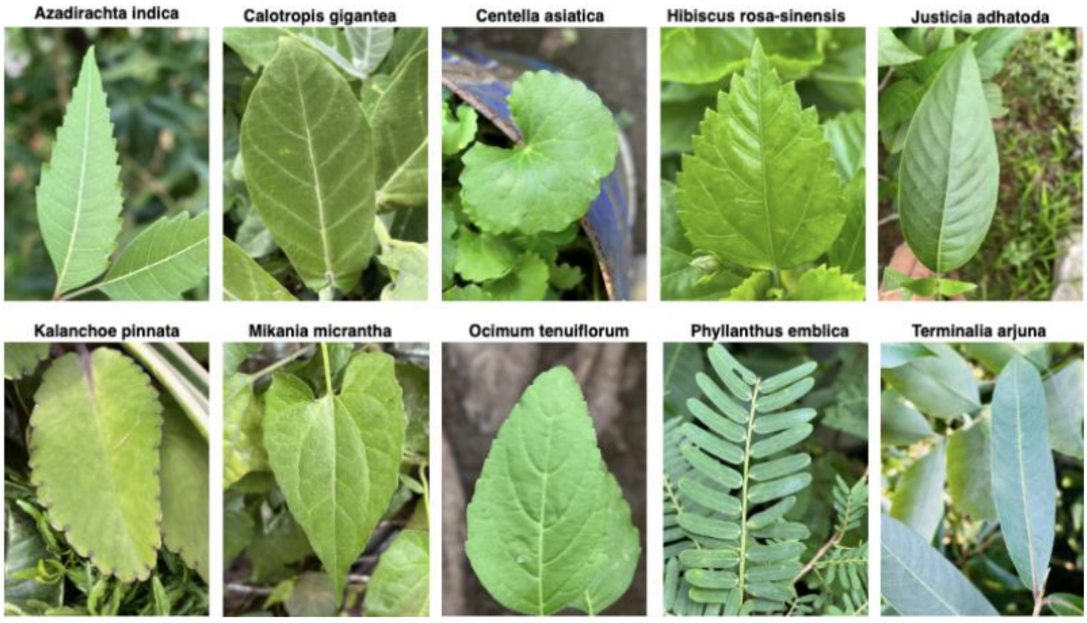

Fig. 1. Sample of data of each of the classes.

Fig. 2. Sample of data before and after denoising.

thresholding. The revised coefficients are utilized to apply the inverse DWT.

Channel merging: The final step involves the merging of the denoised channels back into a single image.

Models

In this research, five models were experimented. VGG16 which is a CNN-based TL model, two ViT models in the form of Compact Convolutional Transformer (CCT) and ConvMixer, and finally proposed two architectures, which are ConvTransNet which is a hybrid of a smaller CNN and a lightweight ViT architecture and a customized CNN model.

CCT: A compact convolutional transformer or CCT[16] is a novel DL architecture with a combination of CNN and transformers. It is useful to capture spatial information using CNNs for local patterns like edges and textures while using transformers for global context and long-range dependencies. This model is especially useful for image-related

Fig. 3. Architecture for the CCT.

tasks, including classification, object detection, and segmentation. Researchers are making these transformer models more versatile so that they can be used for a wider range of tasks. This will help build more powerful and efficient neural networks. Fig. 3 shows the architecture of CCT model.

ConvMixer: To combine geographical locations, the authors used depth wise convolution, and to combine channel locations, they used pointwise convolution.[17] A fundamental concept from earlier research is that MLPs and self-attention are capable of mixing remote spatial locations, meaning they have an infinitely broad receptive field. As such, the distant spatial locations are mixed using convolutions with an extremely large kernel size. Large receptive fields and content-aware behavior are theoretically possible with self-attention and MLPs, but convolution's inductive bias makes it ideal for vision tasks and produces high data efficiency. In addition, they are able to observe the impact of the patch representation itself, which is different from the traditional pyramid-shaped, progressively down-sampling architecture of convolutional net-works, by employing such a standard technique. Fig. 4 shows the architecture of our ConvMixer model.

VGG16: VGG16 stands out as a CNN renowned for its exceptional performance in computer vision.[18] The convolutional layers are stacked in a straightforward manner with small 3 × 3 filters, and max-pooling layers are used to reduce the spatial dimensions. This modification led to a substantial improvement over previous configurations, ultimately expanding the network's depth to include 16 to 19 weight layers, resulting in approximately 138 trainable parameters. The entire architecture of the model is illustrated in Fig. 4.

CNN: The Conv2D layer, the fundamental component of a CNN, is the first layer. It processes the input image through 32 convolutional filters, each measuring 3 by 3 pixels. Each input image must have three color channels (RGB) and be 32 by 32 pixels, according to the input shape argument. By padding the input's edges, the padding "same" option guarantees that the spatial dimensions of the output following convolution match those of the input.

An activation layer called ReLU (Rectified Linear Unit) comes after the convolutional layer. By adding non-linearity to the model, this layer enables it to recognize more intricate patterns in the data. Because

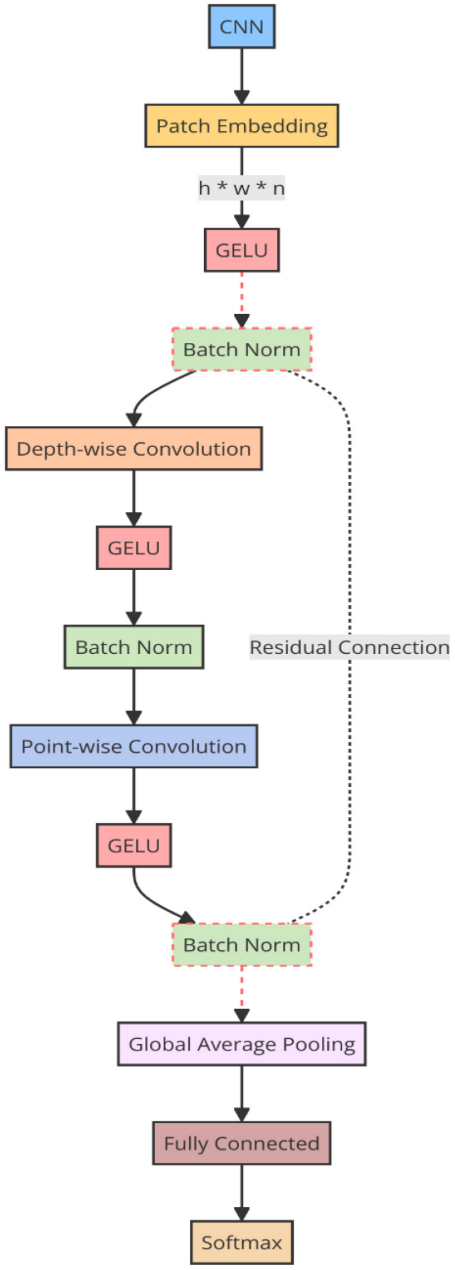

Fig. 4. Architecture for the ConvMixer.

of its effectiveness in minimizing the vanishing gradient problem and its computing efficiency, ReLU is especially well-liked in the field of DL.

Next, the MaxPooling2D layer is employed. Reducing the number of parameters and computations in the network, it downsamples the image along its spatial dimensions (width and height). This layer reduces

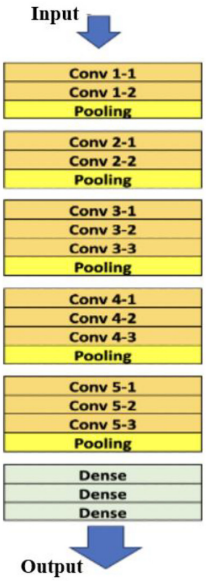

Fig. 5. Architecture for VGG16.

the likelihood of overfitting by taking the greatest value over a 2 × 2 pooling window, hence half the feature map's dimensions. These layers are then repeated by the model: a MaxPooling2D layer, a ReLU activation, and another Conv2D layer. Since many convolutional layers may extract a rich set of features from the input images, CNNs frequently repeat this process. The network may capture more complicated characteristics by increasing the number of filters (64 in this case) in future convolutional layers, see Fig. 5.

Following a number of pooling and convolutional layers, the network becomes a fully connected design. The 2D feature maps are transformed into a 1D vector by the flattened layer. Next, a Dense layer- a kind of fully connected neural network layer with 64 neurons and a ReLU activation function- is applied to this flattened vector. The network can learn non-linear combinations of the high-level information that the convolutional layers have extracted thanks to this layer. Lastly, there is one Denser layer with ten neurons in the output layer. In a multi-class classification issue, this layer is usually followed by a softmax activation function to produce a probability distribution over the classes. Nevertheless, this activation is not mentioned in your model.

Overall, this CNN architecture makes use of convolutional layers to partially process the picture and extract pertinent features, which are subsequently interpreted by dense layers for classification purposes. The entire architecture is highlighted by a pseudocode below.

Algorithm 1 CNN Architecture

1: Initialize Conv2D Layer with 32 filters, kernel size (3, 3), input shape (32, 32, 3), padding "same"
2: Apply ReLU Activation (relu1)
3: Apply MaxPooling2D with size (2, 2)
4: Initialize Conv2D Layer with 64 filters, kernel size (3, 3), padding "same"
5: Apply ReLU Activation (relu2)
6: Apply MaxPooling2D with size (2, 2)
7: Initialize Conv2D Layer with 64 filters, kernel size (3, 3), padding "same"
8: Apply ReLU Activation (relu3)
9: Flatten the output
10: Initialize Dense Layer with 64 units, ReLU activation
11: Initialize Dense Layer with 10 units = 0

ConvTransNet: Similar to the model CCT, our proposed model ConvTransNet attempts to bridge the gap between convolution and attention. But unlike CCT it has a relatively smaller number of parameters and uses a mini architecture of ViT, by this the model aims to harness the strengths of both architectures. Our architecture consists of pooling layers that enable the model to exhibit translation invariance. This means that the extracted features are not affected by mere spatial shifts in the input image. In the transformer block the attention mechanism enables the model to concentrate on certain areas of the image that are more pertinent to a specific task, adjusting the significance of various regions. Additionally, the regularization techniques have used in the form of dropouts and L2 Regularization. By the addition of dropout layers, the architecture increases its robustness and reduces the chance of overfitting. L2 regularization regularizes the convolutional and dense layers with L2 penalties which prevents the network weights from becoming too large, further helping to prevent overfitting. The hybrid nature of the model further helps it to be more scalable by either scaling up or down the model size as per the computation requirements. Moreover, using CNNs only for the initial stages instead of making the whole model ViT-based makes the architecture more efficient. Fig. 6 shows the block diagram of the model.

Results and discussion

The following sections present details of experimental results.

Comparison of model parameters

Table 2 presents a detailed comparison of key parameters for five models used in our research. The

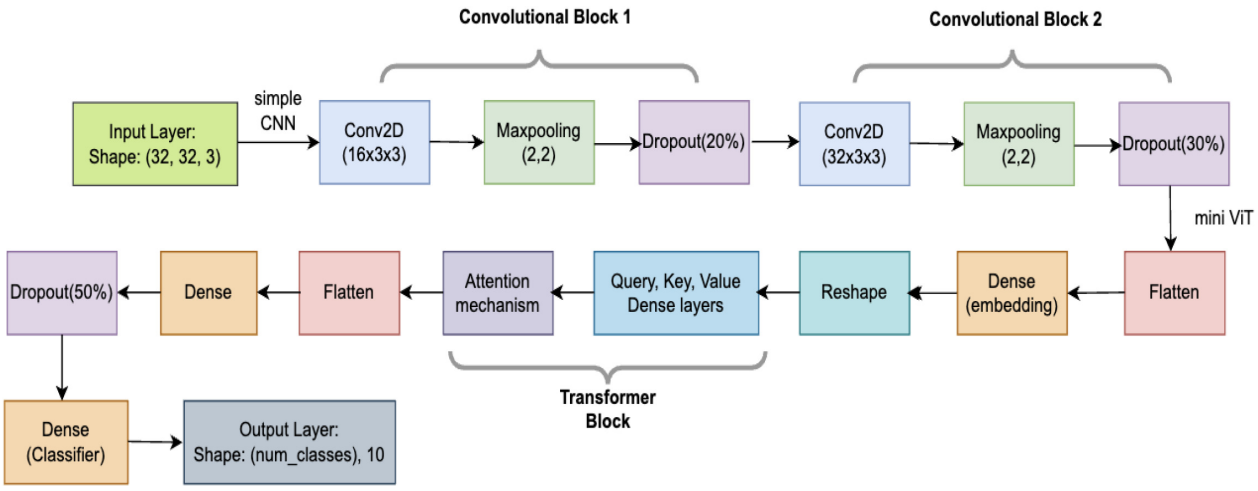

Fig. 6. Architecture for ConvTransNet.

Table 2. Comparison of parameters for the models.

Models	Total Parameters	Trainable Parameters	Non-trainable Parameters	Batch Size	Weight Decay	Learning Rate	Image size	Patch Size	Epochs
CCT	408,139	408,139	0	64	0.0001	0.001	(32,32)	none	30
VGG16	409,357,54	262,210,66	147,146,88	128	none	0.001	(224,224)	none	30
Conv Mixer	602,890	594,186	8704	128	0.0001	0.001	(32,32)	2	30
CNN	319,178	319,178	0	128	none	0.001	(32,32)	none	30
Conv TransNet	341,707	341,707	0	64	none	0.0001	(32,32)	none	30

total number of parameters varies significantly across the models. CCT has 408,139 parameters, VGG16 has the highest with 409, 357, 544, ConvMixer features 602,890, CNN has 319,178, and ConvTransNet comprises 341,707. VGG16 also has the highest number of non-trainable parameters, indicating a reliance on pre-existing features. The distribution of trainable and non-trainable parameters gives an idea of a model's capacity for learning between data and pre-defined features. it also shows that our two proposed models are lightest in terms of parameters compared to other models. Batch size varies for different models. CCT and ConvTransNet used a batch size of 64, and VGG16, ConvMixer, and CNN employed larger batches of 128. Weight decay, a regularization technique preventing overfitting, is used by CCT and ConvMixer with values of 0.0001, while VGG16, CNN, and ConvTransNet do not apply weight decay. Meanwhile, the learning rate varies slightly among all the models.

The dimensions of input images impact a model's feature-capturing ability. CCT processes (32,32) images, VGG16 operates on (224,224), and ConvMixer, CNN, and ConvTransNet handle (32,32) images. ConvMixer was the only one that used a patch size, with a value of 2. This was used to process images by dividing them into smaller patches. A common epoch is used across all the models with a value of 30.

Results

To evaluate the performance of the implemented models, numerous performance evaluation metrics were utilized. These include accuracy curves, confusion matrices, classification reports, and AUC curves. Fig. 7 demonstrates the comparison of each model using an accuracy curve. VGG16 performs the best compared to other models, consistently delivering high performance for validation data. Although its performance increased for 14 epochs, it remained consistent for the remaining epochs. It is observed that training accuracy was around 97% while validation accuracy was close to 93%. In terms of training accuracy, ConvMixer exhibits the highest performance, but its validation accuracy is relatively low, suggesting a possible issue with overfitting. ConvTansnet shows consistency in both training and validation data. While it may not reach the peak accuracy achieved by other models, it avoids overfitting, as observed in the case of ConvMixer. The accuracy of ConvTansnet fluctuates in the first 15 epochs, but there is a steady increase in the last half. The performance of CCT

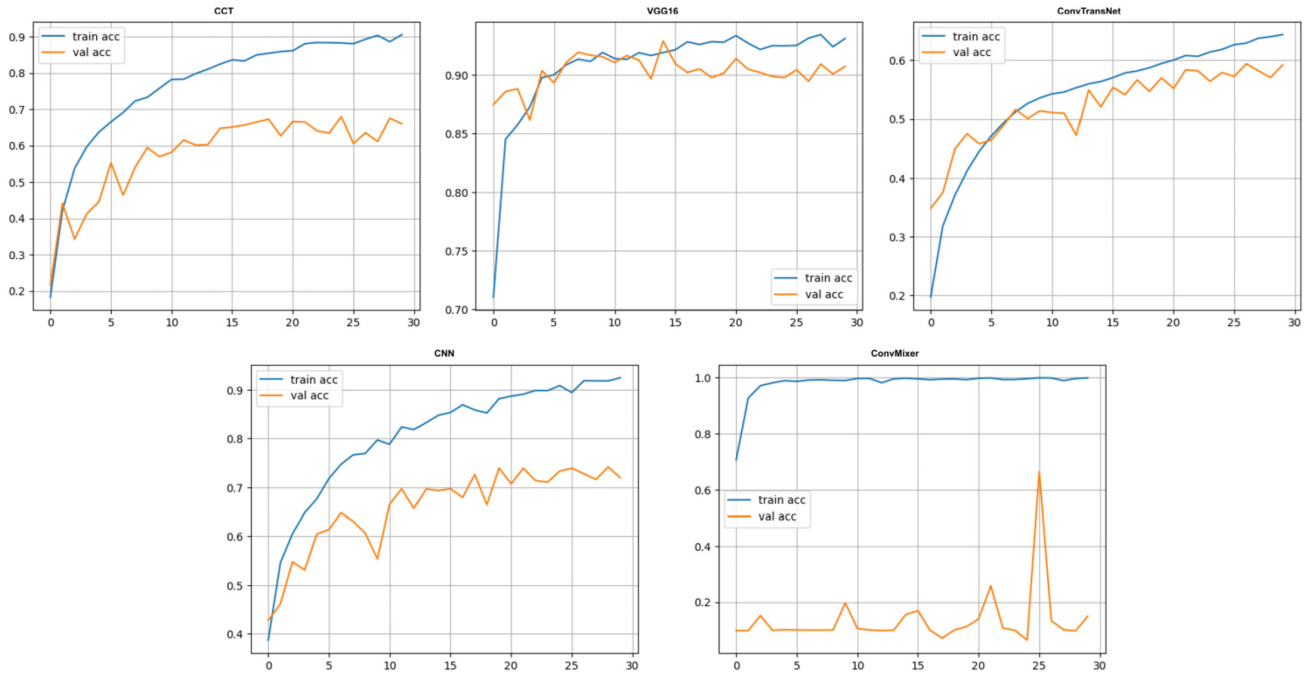

Fig. 7. Accuracy curves for the models.

and CNN is quite similar, both achieving higher accuracy on training data but lower accuracy on validation data. While CCT achieved 91% accuracy for training data, its validation accuracy was below 70%.

Fig. 8 displays the confusion matrices for each model, providing valuable insights into misclassified images. The diagram shows that all models encountered difficulties correctly identifying Azadirachta indica (Neem Tree) images, often misclassifying them as Phyllanthus emblica (Indian gooseberry) tree leaves. Conversely, the models demonstrated consistent success in recognizing Calotropis gigantea images. CCT exhibited a notable challenge by misclassifying 152 Ocimum tenuiflorum (Basil) images as Centella Asiatica (Asiatic pennywort). Furthermore, it struggled significantly with Justicia Adhatoda (Malabar Nut) images, misclassifying 221 instances as Centella asiatica. The confusion matrix for VGG 16 suggests that its performance fell short for various images. ConvTransNet performed reasonably well in image identification but encountered challenges, misclassifying 117 Centella Asiatica images as Ocimum tenuiflorum and 175 Justicia Adhatoda (Malabar Nut) images as Centella Asiatica. CNN demonstrated overall decent performance, with the exception of 87 Kalanchoe Pinnata images, which were misclassified as Phyllanthus Emblica images. However, the performance of ConvMixer was disappointing, as it consistently misclassified a majority of the images as Kalanchoe Pinnata (Cathedral Bells). One possible explanation for this outcome could be overfitting.

The efficiency of each model is assessed using three metrics in our classification report. These are Precision, F1 score, and Recall.

Precision (P) is defined as the ratio of correctly predicted positive results to the total number of positively classified observations. The precision Eq. (2) is given by:

$$P = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{2}$$

Recall (R) is calculated by dividing the number of correctly predicted positive results by the sum of all instances of the original class. The recall Eq. (3) is:

$$R = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{3}$$

The F1-score (F1) is a single score obtained by averaging precision and recall. The F1-score Eq. (4) is:

$$F1 = 2 \times \frac{P \times R}{P + R} \tag{4}$$

Examining the classification report in Fig. 9 reveals that ConvTransNet and CNN outperformed other models in terms of precision, recall, and F1 score. The CCT has achieved the highest precision score

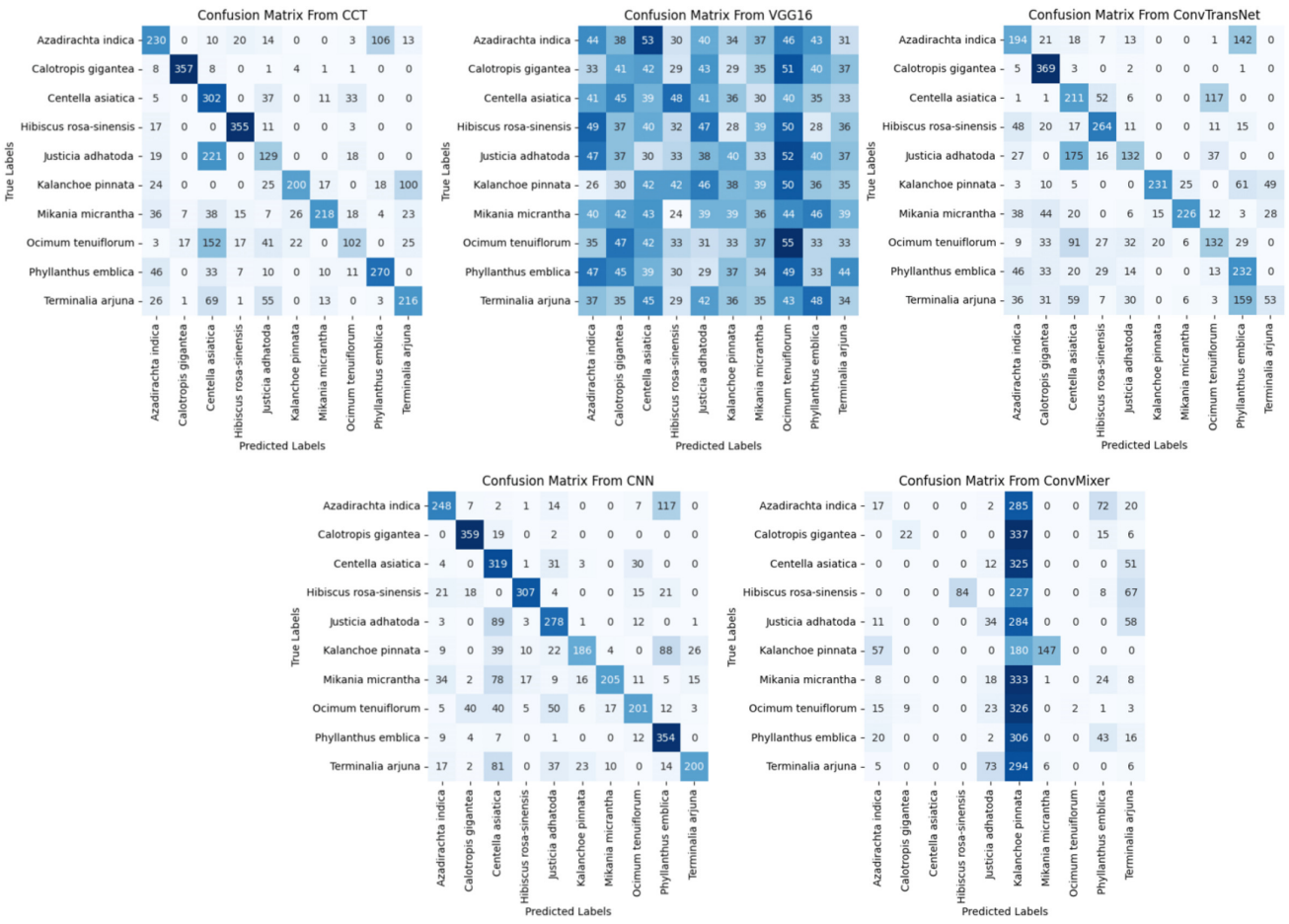

Fig. 8. Confusion matrices for the models.

of 0.93 for Calotropis gigantea, while CNN demonstrated strong precision for Hibiscus Rosa-Sinensis with a score of 0.89. ConvTransNet exhibited the best precision for Kalanchoe Pinnata, achieving a score of 0.87. All three models, CCT, CNN, and ConvTransNet, had good precision scores of 0.81, 0.86, and 0.87, respectively, for Mikania micrantha. On the other hand, VGG16 and ConvMixer displayed subpar performance across all three metrics. ConvTransNet excelled in the recall, achieving the highest score of 0.97 for Calotropis gigantea. CCT and CNN closely followed with recall scores of 0.94 for the same leaf, indicating strong performance across these three models. This also marked the best recall score for CCT and CNN. In terms of F1 score, CNN showed the highest average, followed by CCT and ConvTransNet. Surprisingly, Calotropis gigantea was the plant with the highest F1 scores for CCT (0.94), ConvTransNet (0.78), and CNN (0.88). These three models consistently performed well for Calotropis Gigantea across all three metrics. However, VGG16 and ConvMixer both had relatively lower performance than the other three. In particular, VGG16 failed to surpass a score of 0.15 in any of the three metrics. Despite ConvMixer achieving a relatively good precision score of 0.71 for Calotropis gigantea, its performance in recall and F1 score was also weak.

Fig. 10 illustrates the Area Under Curve (AUC) for each model. VGG16, CCT, and ConvMixer produce average AUC scores of 0.50, 0.79, and 0.50 respectively while our proposed CNN and hybrid model give average AUC scores of 0.83 and 0.74 respectively. CNN got the highest range of AUC scores, ranging from 0.74 to 0.96. Both CCT and ConvTransNet demonstrated consistent performance, with CCT ranging from 0.62 to 0.97, and ConvTransNet ranging from 0.56 to 0.96. Each of these three models performed well when classifying Calotropis Gigantea, achieving scores above 96%. In contrast, both VGG16 and ConvMixer underperformed. VGG16 achieved its highest score of 0.51 for Ocimum Tenuiflorum, while ConvMixer reached 0.61 for Hibiscus Rosa-Sinensis. All models faced challenges when dealing with Ocimum Tenuiflorum, with

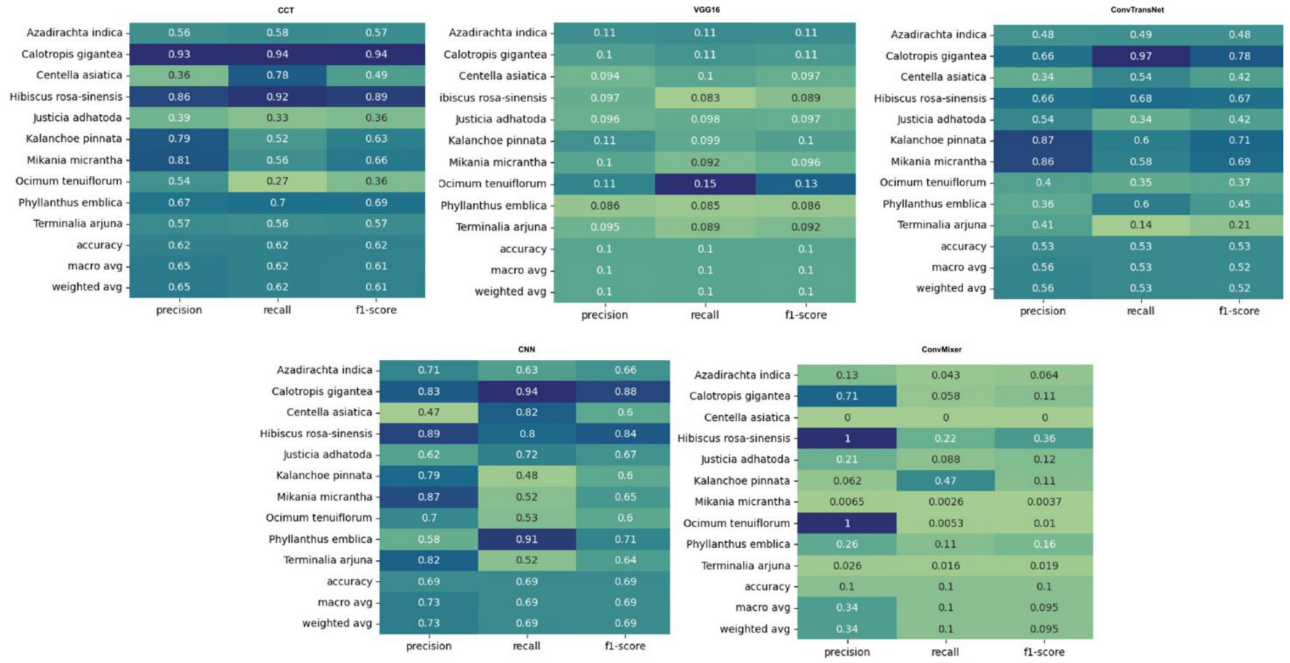

Fig. 9. Classification reports for models.

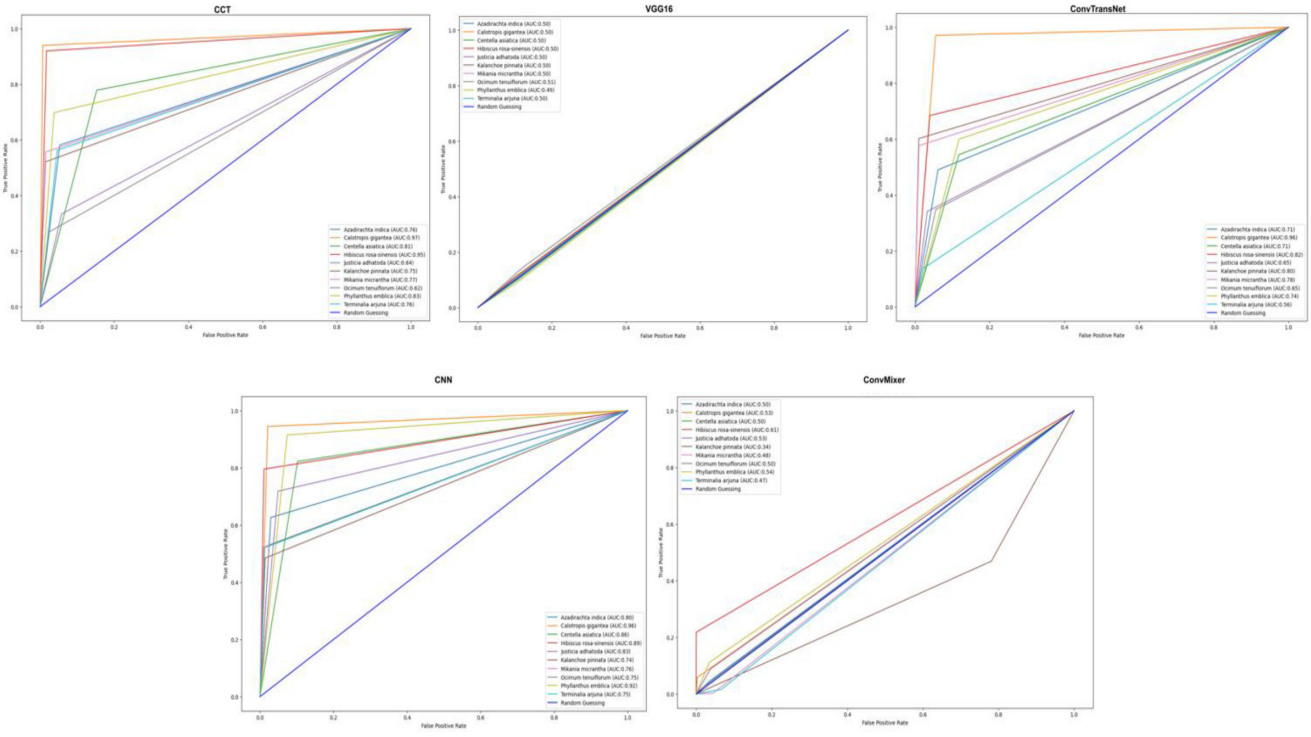

Fig. 10. AUC for the models.

CNN securing the highest score of 0.74 only. One possible reason could be that the images of Ocimum Tenuiflorum may contain complex patterns, variations, or subtle features that are challenging for the models to accurately identify.

Conclusion

In this research, a hybrid denoising algorithm that composed of Wavelet thresholding and Gaussing blurring is presented that ensures both feature prevention

and noise reduction. Additionally, five different DL and ViT-based models are experimented, where the proposed hybrid model of CNN-ViT and improvised CNN had the lowest number of parameters and produced comparable results with the SOTA models, such as VGG16, ConvMixer, and CCT. It is expected that this research will be fruitful for researchers working with similar medicinal data using similar classification tools. In the future we would like to focus on deploying these classifier models into resource constrained edge devices for on field classification of medicinal leaves.

Acknowledgment

This research is funded by Woosong University Academic Research 2024.

Authors' declaration

- Conflicts of Interest: None.
- We hereby confirm that all the Figures and Tables in the manuscript are ours. Furthermore, any Figures and images, that are not ours, have been included with the necessary permission for republication, which is attached to the manuscript.
- No animal studies are present in the manuscript.
- No human studies are present in the manuscript.
- Ethical Clearance: The project was approved by the local ethical committee at Woosong University.

Authors' contribution statement

This work was carried out in collaboration between all authors. S.H. and R.H. wrote the overall paper along with the experimental analysis. J.U. supervised the project, reviewed the manuscript.

References

1. Hassoon IM, Qassir SA, Riyadh M. PDCNN: Framework for potato diseases classification based on feed foreword neural network. Baghdad Sci J. 2021;18(2 (Suppl.)):1012–1012. https://doi.org/10.21123/bsj.2023.9120.
2. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM. 2017 May 24;60(6):84–90. https://doi.org/10.1145/3065386.
3. Jassim OA, Abed MJ, Saied ZH. Indoor/outdoor deep learning based image classification for object recognition applications. Baghdad Sci J. 2023;20(6 (Suppl.)):2540–2540. https://doi.org/10.21123/bsj.2023.8177
4. Abdullah TH, Alizadeh F, Abdullah BH. COVID-19 diagnosis system using SimpNet deep model. Baghdad Sci J. 2022;19:1078–1089. https://doi.org/10.21123/bsj.2022.6074.
5. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint. 2020 Oct 22;1–22. https://doi.org/10.48550/arXiv.2010.11929.
6. Ariful Hassan M, Sydul Islam M, Mehedi Hasan M, Shorif SB, Tarek Habib M, Uddin MS. Medicinal plant recognition from leaf images using deep learning. In Computer Vision and Machine Learning in Agriculture. Singapore: Springer Singapore. 2022 Mar 14;137–154. https://doi.org/10.1007/978-981-16-9991-7_9.
7. Chanyal H, Yadav RK, Saini DK. Classification of medicinal plants leaves using deep learning technique: A review. Int J Int Sys App Eng. 2022 Dec 16;10(4):78–87. https://orcid.org/0000-0001-9678-0125.
8. Naeem S, Ali A, Chesneau C, Tahir MH, Jamal F, Sherwani RA, Ul Hassan M. The classification of medicinal plant leaves based on multispectral and texture feature using machine learning approach. J Agron. 2021 Jan 30;11(2):1–15. https://doi.org/10.3390/agronomy11020263.
9. Kaur M, Bhatia R. Development of an improved tomato leaf disease detection and classification method. In IEEE Conference on Information and Communication Technology. IEEE. 2019;Dec 6:1–5. https://doi.org/10.1109/CICT48419.2019.9066230.
10. Rani L, Devika G, Karegowda AG, Vidya S, Bhat S. Identification of medicinal leaves using state of art deep learning techniques. In IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE). IEEE. 2022;Apr 23:1–5. https://doi.org/10.1109/ICDCECE53908.2022.9792712.
11. Ahmed MR, Fahim MA, Islam AM, Islam S, Shatabda S. DOLG-NeXt: Convolutional neural network with deep orthogonal fusion of local and global features for biomedical image segmentation. J Neurocom. 2023 Aug 14;546:126362. https://doi.org/10.1016/j.neucom.2023.126362.
12. Islam S, Ahmed MR, Islam S, Rishad MM, Ahmed S, Utshow TR, Siam MI. BDMediLeaves: A leaf images dataset for Bangladeshi medicinal plants identification. J Data Brief. 2023 Oct 1;50:109488. https://doi.org/10.1016/j.dib.2023.109488.
13. Chithra K, Santhanam T. Hybrid denoising technique for suppressing Gaussian noise in medical images. IEEE Int Conf PCSI. IEEE. 2017 Sep 21;1460–1463. https://doi.org/10.1109/ICPCSI.2017.8391954.
14. Al Jumah A. Denoising of an image using discrete stationary wavelet transform and various thresholding techniques. JSIP 2013;4:33–41. https://doi.org/10.4236/jsip.2013.41004.
15. Devi TG, Patil N, Rai S, Philipose CS. Gaussian blurring technique for detecting and classifying acute lymphoblastic leukemia cancer cells from microscopic biopsy images. J Life. 2023 Jan 28;13(2):1–12. https://doi.org/10.3390/life13020348.
16. Hassani A, Walton S, Shah N, Abuduweili A, Li J, Shi H. Escaping the big data paradigm with compact transformers. arXiv preprint arXiv:2104.05704. 2021 Apr 12;1–18. https://doi.org/10.48550/arXiv.2104.05704.
17. Trockman A, Kolter JZ. Patches are all you need?. arXiv preprint arXiv. 2022 Jan 24;1–16. https://doi.org/10.48550/arXiv.2201.09792
18. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv. 2014;1409:1556, 1–14. https://doi.org/10.48550/arXiv.1409.1556.

تصنيف أوراق النباتات الطبية باستخدام التعلم العميق ومحولات الرؤية

شهريار حسين[1]، رضبان الحسن[1]، جيا الدين[2]

[1]قسم علوم الحاسب، جامعة جورج ميسون، فيرتاكس، الولايات المتحدة الأمريكية.
[2]قسم الذكاء الاصطناعي والبيانات الضخمة، كلية إنديكوت، جامعة ووسونغ، دايجون، كوريا الجنوبية.

الخلاصة

يعد تحديد أوراق النباتات الطبية أمرًا بالغ الأهمية نظرًا لأن زراعتها وإنتاجها ضروريان لصناعة الأدوية. تبدو العديد من فئات الأوراق الطبية المختلفة متطابقة ولكنها تخدم أغراضًا مختلفة في صناعة الأدوية ولها علاجات مختلفة لأمراض مختلفة. ومن ثم، لا بد من استخدام أساليب آلية وأسرع وتنتج دقة جيدة. لقد تم تدريب النماذج المتطورة لتمييز الفروق الدقيقة بين الأنواع المختلفة من الأوراق، وهو ما يمثل عددًا لا يحصى من العوامل مثل نسيج الورقة، وشكلها، وتغيرات الألوان، والتي غالبًا ما تكون غير محسوسة بالعين البشرية. في هذا البحث، قمنا بتنفيذ نموذج VGG16 القائم على التعلم النقلي (TL) ونماذج محولات الرؤية (ViT) مثل ConvMixer و CCTلتصنيف صور الأوراق الطبية باستخدام مجموعة بيانات مكونة من 38066 صورة للأوراق تحتوي على 10 فئات مختلفة. نقترح نموذجًا مخصصًا للشبكة العصبية التلافيفية (CNN) ونموذج CNN-ViT المختلط، حيث يحتوي كلاهما على عدد منخفض جدًا من المعلمات مقارنة بالنماذج الأخرى بالمقارنة مما يجعلها خفيفة وقادرة على أن تكون أقل تكلفة من الناحية الحسابية. قمنا بمقارنة معلماتها وأدائها مع النماذج المذكورة أعلاه بعد تدريب كل نموذج لمدة 30 حقبة. تنتج VGG16 و CCTو ConvMixerدرجات AUC تبلغ 0.50 و0.79 و0.50 على التوالي لمجموعة البيانات بينما يمنح نموذج CNN والنموذج الهجين المقترح درجات AUC تبلغ 0.83 و0.74 على التوالي. علاوة على ذلك، فإننا نقدم خوارزمية تقليل الضوضاء الهجينة التي تدمج عتبة المويجات والتمويه الغاوسي لتقليل الضوضاء في الصور والحفاظ في نفس الوقت على الجودة الأصلية للصورة.

الكلمات المفتاحية: الشبكة العصبية التلافيفية، التعلم العميق، التمويه الضبابي، تقليل تشويش الصورة، نقل التعلم، محول الرؤية، عتبة المويجات، VGG16، CCT، ConvMixer، CNN-ViT.