Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 234 (2024) 247–254

Procedia Computer Science

www.elsevier.com/locate/procedia

Seventh Information Systems International Conference (ISICO 2023)

Medicinal Plants Identification Using Federated Deep Learning

Khanh Le Dinh Viet[a], Khiem Le Ha[a], Trung Nguyen Quoc[a], Vinh Truong Hoang[b]

[a] *Department of Information Technology, FPT University, Ho Chi Minh City, Vietnam*
[b] *Faculty of Information Technology, Ho Chi Minh City Open University, Vietnam*

Abstract

Over the years, scientists have discovered bioactive chemicals in many of the plants that have been traditionally utilized as medicinal medicines. However, identifying plant species based on their physical characteristics can be difficult, and misidentification can have severe consequences, such as the use of the incorrect plant as a medicine. With the advent of machine learning techniques such as deep learning and federated learning, it is now possible to develop automated systems for the precise image-based classification of medicinal plants. Nevertheless, medicinal plant classification using deep learning techniques typically requires a large amount of data, which can be challenging to acquire and manage due to privacy concerns, data ownership, and geographic reasons. Federated learning provides a solution to this issue by enabling the training of a shared model on multiple devices without requiring centralized data storage. In this work, we assess and optimize the federared learning framework using two federated learning approaches, FedAvg and FedProx, and four state-of-the-art deep learning networks for the job of categorizing medicinal plants by distributing the original training set into two forms, IID and Non-IID. Ultimately, the accuracy of the optimal federated learning system is improved by 5.65% and 14.84% over the baseline on IID data and Non-IID data, respectively. Furthermore, the study brings up a new difficult arena for the task of classifying medicinal plants using Non-IID training data.

Keywords: Medicinal Plants; Federated Learning; Classification; Deep Learning

1. Introduction

Since ancient times, medicinal plants have been utilized to treat a wide range of ailments and diseases. Although the use of medicinal plants has decreased as modern medicine has developed, their significance in traditional and alternative medicine has not diminished. The identification and use of medicinal plants in traditional medicine, as well as their preservation and protection in the wild, depend on their classification. However, manual observation and study of plant traits used in conventional methods of classification can be time-consuming and prone to inaccuracy.

E-mail address: {khanhldvse150118, khiemlhse150198, trungnq46}@fpt.edu.vn; Corresponding author: vinh.th@ou.edu.vn

10.1016/j.procs.2024.02.171

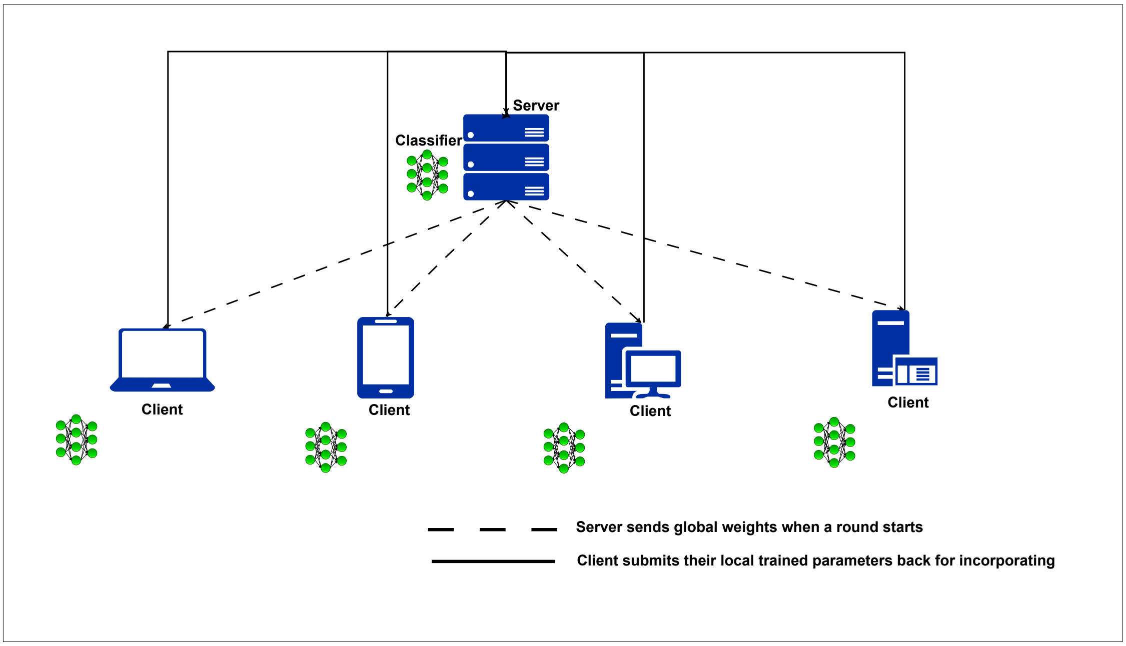


Fig. 1: The illustration of federated learning system.

The accuracy and speed of medicinal plant classification have recently showed significant promise thanks to recent developments in machine learning, particularly deep learning. However, the availability and caliber of training data have a significant impact on how well machine learning models perform. Given that these plants are frequently found in secluded and difficult-to-reach places, gathering big and diverse datasets for the classification of medicinal plants can be a considerable difficulty.

A new machine learning paradigm called federated learning [21] tackles the difficulties associated with training models on distributed and decentralized data. Without sharing their data with a central server, several clients, each with their own dataset, work together to build a shared machine learning model in federated learning. This method provides various benefits, including higher scalability, improved data privacy, and less communication expenses.

In this study, we investigate the use of federated learning in the identification of medicinal plants. By utilizing two federated learning algorithms, FedAvg [10] and FedProx [7], we specifically study the efficacy of federated learning in training deep learning models on decentralized and distributed datasets of medicinal plant photos. In a restricted number of communication rounds, we also assess the effects of several model architectures and hyper-parameters on the accuracy of classification results and compare how well federated learning performs on two data distribution methodologies.

The remainder of the essay is structured as follows. In Section 2, we examine relevant research on federated learning and the classification of medicinal plants. We outline our federated learning algorithm and the used dataset for classifying medicinal plants in Section 3 of this paper. We give experimental findings and contrast the effectiveness of federated learning with different models and hyper-parameters in Section 4. We wrap up the ramifications of our findings in Section 5 and suggest ideas for new research trajectories.

2. Related Works

We employ federated machine learning algorithms to a medicinal plant dataset in order to observe the resulting effect. In almost previous research, the majority of works in federated learning utilize canonical datasets such as MNIST, CIFAR-10, and their variants, which contributes to a negative bias. In particular, this lack of generalization can prevent external individuals or organizations from utilizing federated learning in their products or service solutions, as they lack solid evidence that all research conclusions are independent of domain data or classifier architectures.

Since the early 2000s, a number of concepts regarding the partitioning of computing tasks have been explored. On structured perceptron, iterative parameter mixing implements the concept that most closely resembles how the federated learning technique is constructed [9]. In addition, some publications investigate distributed optimization methods [2, 24]. These works focus solely on reducing complexity and maximizing available hardware resources in order to accelerate the learning process (data is gathered at one location). Federated Learning is the result of integrating previous works in response to the need for a model that enables the security and use of massive data on end devices [10]. In this new context, there are inherent challenges (we follow these challenges in directing our experiments, which will be presented in greater detail in subsequent sections): (1) privacy concerns; (2) the disparity in client data regarding size, feature space, and data distribution; (3) different hardware specifications; and (4) convergence assurance when compared to a centralized situation.

FedAvg [10] illustrates that client diversity is the most critical factor affecting our performance. Some recent investigations have attempted to address this issue, but they are not exhaustive. In spite of the fact that FedAvg is an empirical technique that functions well in specific contexts under the condition that hyper-parameters are properly tuned, more recent theoretical works support the robustness of this method [22, 23]. However, the authors presume that every device participates in each round of the process and that the used solver is typically predefined (either SGD or GD). Exposing a client's data to other clients or to the coordinator is a strategy for addressing the heterogeneous issue. Nonetheless, this imposes a significant stress on network bandwidth (especially in environments with expensive network connections) and simultaneously violates privacy standards. FedProx [7] provides a more comprehensive theoretical framework for handling heterogeneous data than previous works. Through a mechanism that permits some clients to submit their truncated parameters, the authors also accommodate for the disparity in computational capabilities between clients.

Numerous researchers are drawn to the identification of medicinal plants due to its widespread applications in both the medicinal community and industry. Regarding datasets, the majority of works rely on their own self-collected datasets, which typically offer distinct properties because each nation has a unique biologic ecosystem. This complicates the process of comparing attained results for the purpose of leveraging existing models, as collectors utilize various lighting, perspectives, sizes, and backgrounds when taking photographs. Currently employed leaf recognition datasets include Flavia [20], Swedish Leaf [16], ICL [19], Leafsnap [5]. The majority of the images were captured in controlled environments, and each represents a distinct group of plants. Evidently, identifying a single leaf in indoor conditions is a far away from identifying a plant in an outdoor setting captured by a handheld device. Several papers on medicinal plants from India and Southeast Asia have been proposed with different datasets [14, 6, 1, 18, 11]. Leaf detection could be utilized to improve overall performance; it requires image preprocessing, image enhancement, or even localization and segmentation. Gao and Lin [3] employ OTSU, an effective segmentation algorithm, to increase their accuracy to 99.9%. Typical feature extractors include HOG, LBP, the transform technique, and deep learning models.

VNPlant-200 [13] is regarded as the first publicly available actual dataset on Vietnamese herbs. The dataset includes 20,000 images of 200 species, with 12,000 used for training and the remainder for testing. The images are quite challenging due to the fact that it stimulates outdoor perspective with a variety of noise objects and varying points of view. Using SIFT and SURF feature extractors in conjunction with Random Forest classifier yields modest results as a baseline [13]. In [12], the author adopted multiple CNN classifiers, including VGG, Inception V3, MobileNetv2, Resnet50, DenseNet, and Xception, which significantly improves accuracy. Another group extends their experiments to numerous state-of-the-art classification backbone models and provides a tuning framework for hyper parameter. In addition, they conduct time-efficient comparisons in their task.

3. Methods

3.1. Dataset

The VNPlant-200 dataset [13] is utilized in this study to examine how well the federated learning architecture performs when classifying medicinal plants. Figure 2 demonstrates several medicinal plant samples of VNPlant-200.

Fig. 2: The demonstration of VNPlant-200 dataset.

With a percentage of 50%, 10%, and 40%, respectively, the original dataset is separated into training, validation, and testing sets. Following that, the training images are dispersed to 10 clients using either the independent and identical distribution (IID) method or the non-independent and identical distribution (Non-IID) method. In the IID technique, clients are randomly assigned training data, resulting in data that is distributed similarly across all clients. Instead, the Non-IID technique sorts medicinal plants according to their labels before seeding the data into clients in the appropriate sequence. When using federated learning, the second strategy might reflect a heterogeneous property of decentralize data in the real world. The process of identifying medicinal plants would be more difficult than earlier similar efforts due to the diversity distribution among each client, and this would provide a new avenue for classification optimization.

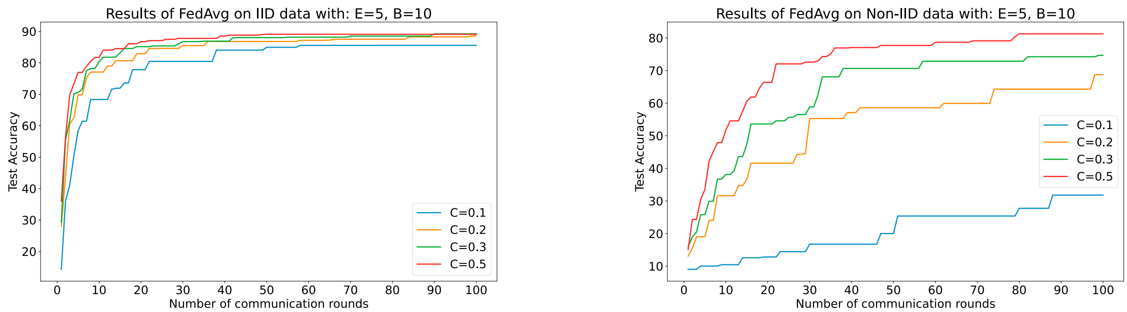
3.2. Federated Learning Frameworks For Classification

The suggested medicinal plant identification frameworks utilizing federated learning include two key components: classifiers and federated learning algorithms. The demonstration of our federated learning systems is shown in Figure 1. Four contemporary deep learning architectures, namely VGG16 [15], ResNet50 [4], ConvNext [8], and MaxVit [17], are incorporated into the framework to enhance identification performance for the classification models. In the context of federated learning, at each round of communication, the classifier parameters of trained clients are sent to the central server, which then employs federated algorithms as an aggregation method for handling clients' parameters in order to update the global model. Figure 1 illustrates how a federated learning system works.

FedAvg [10] is based on a basic but effective concept. A C portion of clients would participate in the training procedure during each communication round. The located data would be looped through E epochs and B batch size for each client. After local tasks have been completed, the weights of each classifier will be averaged to update all client models. However, arbitrarily averaging the model weights could result in an unstable training process if the difference between training data from each communication round is significant. The FedProx [7] algorithm may improve classification performance through a more stable coverage process by incorporating proximal terms into loss functions in order to solve this issue.

4. Experimental Results

Many experiments are conducted to optimize federated learning framework for the best medicinal plants classification performance in a fixed number of communication rounds. To optimize federated learning framework for medicinal plant categorization in a fixed number of communication rounds, many experiments are done. All of the experiments are executed on Google Colab and require 1000 computing units, which is equivalent to 500 hours of training. Furthermore, while PyTorch is utilized to configure classification models, the decentralized training script is built in Python without additional library assistance.



(a) IID data (b) Non-IID data

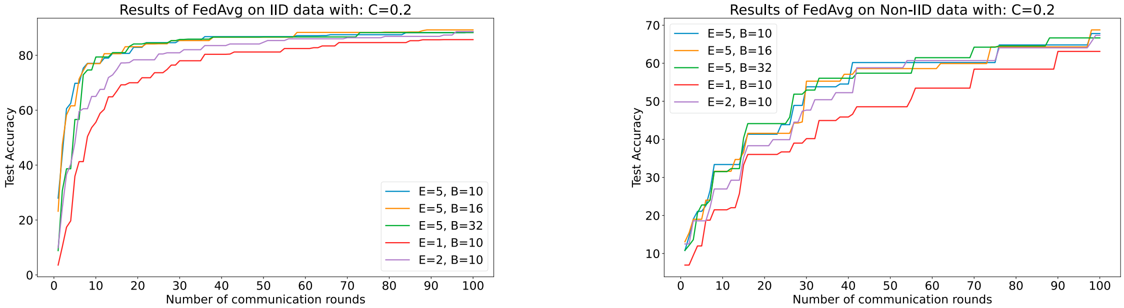
Fig. 3: Classification results of VGG16, FedAvg with B = 10, E = 5 and different C

C	IID				Non-IID			
	10	20	50	100	10	20	50	100
0.1	68.34	77.84	84.95	85.56	10.44	12.81	20.00	31.80
0.2	77.08	82.90	86.80	88.56	33.39	41.35	60.19	67.81
0.3	80.34	85.15	88.04	89.24	38.13	53.61	70.65	74.68
0.5	81.71	86.76	89.09	89.09	51.73	66.40	77.71	81.28

Table 1: Classification results of VGG16, FedAvg with $B = 10$, $E = 5$ and different C

In the initial phase, the objective of tuning experiments is to determine appropriate values for C, B, and E using the baseline framework of VGG16 and FedAVg as a classifier and federated learning algorithm, respectively. Table 1 and Figure 3 displays the framework's medicinal plant identification using VGG16 and FedAvg with $B = 10$, $E = 5$, and increasing C values after 10, 20, 50, and 100 communication cycles. hen more clients are involved in each training round at once, the categorization performance improves. In addition, the extent of influence between IID and Non-IID data differs. Specifically, on Non-IID data, the classification results improve more than IID data on each increment value of C, which can be explained by the unique data distribution of each Non-IID client, but the data distribution of IID clients is similar to the worldwide distribution. For the sake of computation, subsequent experiments fix C to 0.2 and tune additional variables such as B, E, and the classifier.

Table 2 and Figure 4 demonstrates the classification performance of the proposed framework with varying values of B and E. When increasing the batch size from $B = 10$ to $B = 16$, the highest accuracy for inspected rounds also



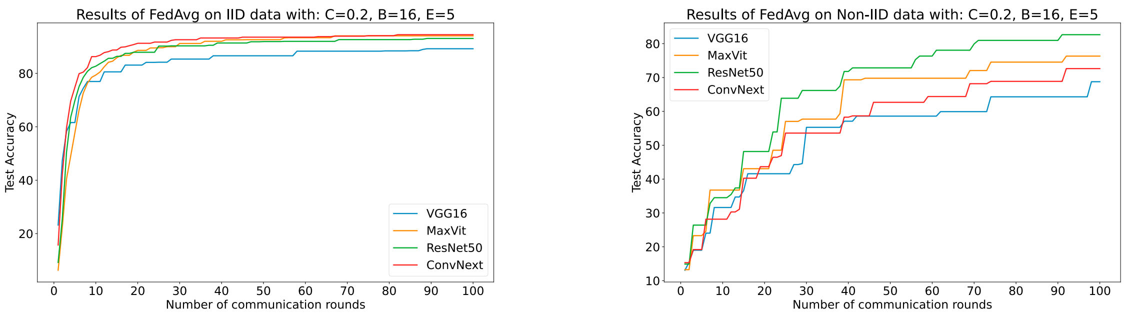
(a) IID data (b) Non-IID data

Fig. 4: Tuning results of VGG16 and FedAvg with C = 0.2

E	B	IID				Non-IID			
		10	20	50	100	10	20	50	100
5	10	77.08	82.90	86.80	88.56	33.39	41.35	60.19	67.81
5	16	78.48	83.41	87.59	88.93	31.61	41.59	58.60	68.76
5	32	79.38	84.13	86.69	88.28	31.51	44.14	57.39	66.66
1	10	55.60	70.00	81.20	85.71	21.49	36.04	48.56	63.11
2	10	65.03	78.36	84.75	88.64	26.96	38.33	58.53	67.34

Table 2: Tuning results of VGG16 and FedAvg with $C = 0.2$

improves substantially. Following 100 rounds, the accuracy of IID data grew by 0.37%, from 88.56% to 88.93%, while the accuracy of Non-IID data climbed by 0.95%, from 67.81% to 68.76%. However, consistently increasing B to 32 does not result in a significant improvement comparable to $B = 16$. Thus, $B = 16$ would be an optimal value of B in the federated learning framework for medicinal plant classification. To avoid over-fitting of the local model during training progress, small values of epoch E are used in the experiments. For $E = 1$ and $E = 2$, there is no improvement in the training stage for either IID or Non-IID data, so the optimal number of epochs is $E = 5$.



(a) IID data (b) Non-IID data

Fig. 5: Classification results on different models using FedAvg with $C = 0.2$, $B = 16$, and $E = 5$

After determining the most suitable hyper-parameters for the framework, a number of contemporary deep learning networks are used as classifiers to determine which could yield the highest accuracy. These experimental outcomes are displayed in Table 3 and Figure 5. Using ConvNext as a classification model considerably increases the final accuracy of IID results from 88.93% to 94.51%. In the meantime, after 100 communication cycles, ResNet50 is the best model for classifying Non-IID medicinal plant data with 82.65% accuracy, a 13.92% improvement over VGG16's 68.76% accuracy.

Model	IID				Non-IID			
	10	20	50	100	10	20	50	100
VGG16	78.48	83.41	87.59	88.93	31.61	41.59	58.60	68.76
ConvNext	86.40	91.29	93.59	94.51	30.86	48.15	68.11	73.09
ResNet50	82.73	87.93	91.94	93.10	34.51	48.15	72.85	82.65
MaxVit	79.41	88.64	92.65	94.01	36.76	43.08	69.79	76.33

Table 3: Classification results on different models using FedAvg with $C = 0.2$, $B = 16$, and $E = 5$

Muy	Non-IID			
	10	20	50	100
0	34.51	48.15	72.85	82.65
1	35.48	50.00	73.23	82.95
0.1	34.76	48.48	73.13	82.37
0.01	34.34	50.93	73.10	82.71
0.001	35.24	49.60	73.81	82.41

Table 4: Medical plants classification results with FedProx, ResNet50, $C = 0.2$, $B = 16$, and $E = 5$

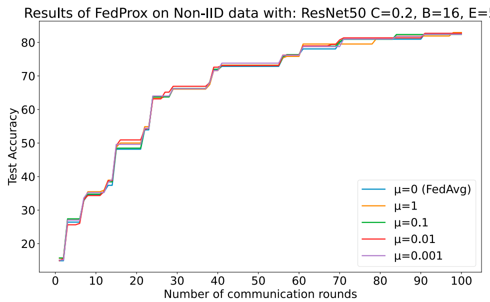


Fig. 6: Medical plants classification results with FedProx, ResNet50, $C = 0.2$, $B = 16$, and $E = 5$

Despite the suggested framework achieves excellent performance with FedAvg on IID data, with a peak of 94.51% after 100 communication rounds, the task of classifying Non-IID remains difficult, with a final accuracy of just 82.65%. Individual clients' disparate data distributions slow down the convergence of classification models and hinder global models from correctly representing the distribution of data. FedProx is therefore anticipated to maintain the training process' stability by including a proximal term in the loss function, which may be managed by changing the value of μ. The results of the federated learning framework utilizing ResNet50 and FedProx with diverse μ values are presented in Table 4 and Figure 6. In comparison to the findings of FedAvg (μ=0), all FedProx tests produce superior results, reaching a peak at $\mu = 1$ with 0.38% improved accuracy after 50 rounds and 0.30% improved accuracy after 100 rounds.

5. Conclusion

In this work, the usefulness of federated learning for medicinal plant classification was investigated utilizing both IID and Non-IID data. FedAvg and FedProx algorithms were utilized to train a deep learning classifier on a large dataset of medicinal plant images that were distributed across multiple participating devices without the need to share data. The performance of our federated learning system was enhanced by adjusting hyper-parameters including the batch size B, number of epochs E, classifier model, and control value of proximal term *mu*. Additionally, we have shown how FedProx outperforms FedAvg in terms of accelerating convergence and strengthening the training process, especially apparent for Non-IID data. In the end, after 100 communication rounds, the fantastic performance of the ideal framework helped enhance 5.95% accuracy on IID data and 14.84% accuracy on Non-IID data compared to the baseline design. Moreover, we discovered that the efficacy of our federated learning system with Non-IID data was inferior to that with IID data. The performance of the federated learning approach may suffer as a result of the dissemination of Non-IID data, according to this.

Overall, the findings of this study indicate that federated learning is a promising approach for the classification of medicinal plants and other applications where privacy and data security are crucial. Nonetheless, the efficacy of the federated learning approach may be impacted by the data distribution, particularly when Non-IID data are involved. Future research could investigate the use of other, more complex federated learning algorithms and further hyper-parameter optimization to enhance the system's efficacy on Non-IID data.

References

[1] Arun, C.H., Durairaj, D.C., 2017. Identifying medicinal plant leaves using textures and optimal colour spaces channel. Jurnal Ilmu Komputer dan Informasi 10, 19. doi:10.21609/jiki.v10i1.405.

[2] Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Mao, M., aurelio Ranzato, M., Senior, A., Tucker, P., Yang, K., Le, Q., Ng, A., 2012. Large scale distributed deep networks, Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2012/file/6aca97005c68f1206823815f66102863-Paper.pdf.

[3] Gao, L., Lin, X., 2018. A method for accurately segmenting images of medicinal plant leaves with complex backgrounds. Computers and Electronics in Agriculture 155, 426–445. doi:10.1016/j.compag.2018.10.020.

[4] He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition .

[5] Kumar, N., Belhumeur, P.N., Biswas, A., Jacobs, D.W., Kress, W.J., Lopez, I.C., Soares, J.V.B., 2012. Leafsnap: A Computer Vision System for Automatic Plant Species Identification. pp. 502–516. doi:10.1007/978-3-642-33709-3_36.

[6] Le, T.L., Tran, D.T., Hoang, V.N., 2014. Fully automatic leaf-based plant identification, application for vietnamese medicinal plant search, ACM Press. pp. 146–154. doi:10.1145/2676585.2676592.

[7] Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V., 2018. Federated optimization in heterogeneous networks .

[8] Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A convnet for the 2020s, IEEE. pp. 11966–11976. doi:10.1109/CVPR52688.2022.01167.

[9] McDonald, R., Hall, K., Mann, G., 2010. Distributed training strategies for the structured perceptron, in: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Los Angeles, California. pp. 456–464. URL: https://aclanthology.org/N10-1069.

[10] McMahan, H.B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A., 2016. Communication-efficient learning of deep networks from decentralized data URL: http://arxiv.org/abs/1602.05629.

[11] Minh, T.D., Minh, T.T., Quoc, T.N., Hoang, V.T., 2023. Features Extraction Based on Sota Models for Medicinal Plant Images Recognition. pp. 465–473. doi:10.1007/978-3-031-27524-1_44.

[12] Quoc, T.N., Hoang, V.T., 2020. Medicinal plant identification in the wild by using cnn, IEEE. pp. 25–29. doi:10.1109/ICTC49870.2020.9289480.

[13] Quoc, T.N., Hoang, V.T., 2021. VNPlant-200 – A Public and Large-Scale of Vietnamese Medicinal Plant Images Dataset. pp. 406–411. doi:10.1007/978-3-030-49264-9_37.

[14] Sainin, M.S., Ghazali, T.K., Alfred, R., 2012. Malaysian Medicinal Plant Leaf Shape Identification and Classification. Master's thesis. URL: http://www.kmice.cms.net.my/.

[15] Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition .

[16] Söderkvist, O.J.O., 2001. Computer Vision Classification of Leaves from Swedish Trees. Master's thesis.

[17] Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., Li, Y., 2022. Maxvit: Multi-axis vision transformer .

[18] Vo, A.H., Dang, H.T., Nguyen, B.T., Pham, V.H., 2019. Vietnamese herbal plant recognition using deep convolutional features. International Journal of Machine Learning and Computing 9, 363–367. doi:10.18178/ijmlc.2019.9.3.811.

[19] Wang, B., Gao, Y., Sun, C., Blumenstein, M., Salle, J.L., 2017. Can walking and measuring along chord bunches better describe leaf shapes?, IEEE. pp. 2047–2056. doi:10.1109/CVPR.2017.221.

[20] Wu, S.G., Bao, F.S., Xu, E.Y., Wang, Y.X., Chang, Y.F., Xiang, Q.L., 2007. A leaf recognition algorithm for plant classification using probabilistic neural network, IEEE. pp. 11–16. doi:10.1109/ISSPIT.2007.4458016.

[21] Yang, Q., Liu, Y., Chen, T., Tong, Y., 2019. Federated machine learning. ACM Transactions on Intelligent Systems and Technology 10, 1–19. doi:10.1145/3298981.

[22] Yu, H., Jin, R., Yang, S., 2019a. On the linear speedup analysis of communication efficient momentum sgd for distributed non-convex optimization .

[23] Yu, H., Yang, S., Zhu, S., 2019b. Parallel restarted sgd with faster convergence and less communication: Demystifying why model averaging works for deep learning. Proceedings of the AAAI Conference on Artificial Intelligence 33, 5693–5700. doi:10.1609/aaai.v33i01.33015693.

[24] Zhang, S., Choromanska, A.E., LeCun, Y., 2015. Deep learning with elastic averaging sgd, Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2015/file/d18f655c3fce66ca401d5f38b48c89af-Paper.pdf.