Data in Brief 54 (2024) 110281

Contents lists available at ScienceDirect

Data in Brief

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

Invited Data Manuscript

Mulberry leaf dataset for image classification task

Thipwimon Choompookham[a], Emmanuel Okafor[b], Olarik Surinta[c,*]

[a] *Faculty of Information Technology, Rajabhat Maha Sarakham University, Mahasarakham 44000, Thailand*
[b] *SDAIA-KFUPM Joint Research Center for Artificial Intelligence, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia*
[c] *Multi-agent Intelligent Simulation, Laboratory (MISL) Research Unit, Department of Information Technology, Faculty of Informatics, Mahasarakham University, Mahasarakham, 44150 Thailand*

ARTICLE INFO

Article history:
Received 20 November 2023
Revised 3 February 2024
Accepted 26 February 2024
Available online 1 March 2024

Dataset link: Mulberry leaf dataset (Original data)

Keywords:
Mulberry leaf image
Image recognition
Image classification
Deep learning
Ensemble learning

ABSTRACT

This manuscript presents a mulberry leaf dataset collected from five provinces within three regions in Thailand. The dataset contains ten categories of mulberry leaves. We proposed this dataset due to the challenges of classifying leaf images taken in natural environments arising from high inter-class similarity and variations in illumination and background conditions (multiple leaves from a mulberry tree and shadows appearing in the leaf images). We highlight that our research team recorded mulberry leaves independently from various perspectives during our data acquisition using multiple camera types. The mulberry leaf dataset can serve as vital input data passed to computer vision algorithms (conventional deep learning and vision transformer algorithms) for creating image recognition systems. The dataset will allow other researchers to propose novel computer vision techniques to approach mulberry recognition challenges.

* Corresponding author.
E-mail address: olarik.s@msu.ac.th (O. Surinta).
Social media: @mrolarik (O. Surinta)

https://doi.org/10.1016/j.dib.2024.110281

Specifications Table

Subject	Computer Vision, Computer Science
Specific subject area	Computer Vision, Image Processing, Image Classification, and Deep Learning
Type of data	Image
Data collection	We collected the mulberry leaf cultivars from three regions of Thailand (northern, central, and northeast) that included five provinces in total (Chiang Mai, Phitsanulok, Nakhon Ratchasima, Burriram, and Mahasarakham). DSLR and phone cameras were used to take images with different perspectives from ten mulberry leaf cultivars recorded in the natural environments with different perspectives. The mulberry leaf dataset includes 5,262 images of 10 mulberry leaf cultivars: King Red, King White, Taiwan Maechor, Taiwan Strawberry, Black Austurkey, Black Australia, Chiang Mai 60, Buriram 60, Kamphaeng Saen 42, and Mixed Chiang Mai 60+Buriram 60.
Data source location	The mulberry leaf dataset was collected from three regions of Thailand: Northern (Chiang Mai), Central (Phitsanulok), and Northeast (Nakhon Ratchasima, Burriram, and Mahasarakham).
Data accessibility	Mulberry leaf dataset Repository name: Mendeley Data Data identification number: 10.17632/ds45yy9jrc.3 Direct URL to data: https://data.mendeley.com/datasets/ds45yy9jrc/3
Related research article	Chompookham, T. & Surinta, O. (2021). Ensemble methods with deep convolutional neural networks for plant leaf recognition. ICIC Express Letters, 15(6), 553-565. DOI: 10.24507/icicel.15.06.553

1. Value of the Data

- The mulberry leaf dataset offers a set of challenging images with spatial variations of mulberry leaves taken in natural environments from different perspectives. Researchers will find the dataset interesting due to one image containing multiple mulberry leaves and shadow contextual information in the photos.
- The mulberry leaf dataset is a publicly available annotated dataset containing 5,262 samples of mulberry leaf images with a diversity of 10 classes. One of the central goals of making this dataset publicly available is to aid researchers in the field of computer vision in proposing emerging learning algorithms capable of automatically classifying images with different contours and spatial representations of mulberry leaves with high inter-class similarity and illumination variations.
- Computer vision and image processing researchers can considerably benefit from the mulberry leaf dataset, enabling them to propose a novel deep learning framework to approach image recognition tasks.

2. Background

Numerous plant image datasets are available for image classification tasks; for example, Folio, LeafSnap, and Swedish Leaf datasets [1–3] have previously been used for performing plant species identification, and the economic plant [4,5] and PlantVillage [6] datasets were gathered to classify plant leaf with diverse disease conditions. The PlantVillage dataset has over 50,000 plant leaf images of 14 plant species and is categorized into 38 classes, including healthy crops and plant leaves containing varying disease conditions. Furthermore, the economic plant dataset contains healthy and 12 distinct unhealthy plant leaves. However, the above plant leaf image datasets exclusively focus on each leaf on a single-color background, and no other objects or plant leaves are visible in the image.

We propose a mulberry leaf dataset containing ten cultivars of mulberry leaves. This dataset is unique and different from the previously described datasets above due to the following qualities. All the curated samples of the leaf images were taken in natural environments from different angular perspectives due to the orientations of the cameras, and the leave images have varying background conditions such as natural elements (soil, grass, and mulberry trees) and illumination variations (shadows and contrasting views of the leaf image samples). Some previous studies have investigated the utility of deep learning methods to classify mulberry leaves [7–10]. The value of making this dataset publicly available presents the potential to develop emerging computer vision techniques (vision transformers) to surpass the state-of-the-art performance of this dataset. Additionally, the proposed dataset allows the research community to explore novel data augmentation algorithms to tackle or curb class imbalance in the existing dataset.

3. Data Description

We curated the mulberry leaf dataset in 2019 and 2020. It contains 5,262 images categorized into ten classes existing in diverse environmental conditions. No seasonal considerations were factored in during the data collection; however, all data was collected on sunny days. There is no existence of external plants in the samples of the mulberry leaves, but there are some natural backgrounds of the brown soil, grass, and the mulberry tree, which exist in small proportion relative to the mulberry leaves. The image format of the mulberry leaf dataset exists in JPEG format, and we resize the images to a uniform image resolution size of 224 × 224 pixels. The researcher captured mulberry leaf images from various regions and provinces in Thailand for three months. Further, the dataset was annotated by a domain expert responsible for classifying each mulberry leaf into its respective classes or categories. The leaves with similar features or properties were stored in a specific folder (class), resulting in ten possible classes.

Samples of the mulberry leaf images are shown in Fig. 1. The number of the mulberry leaf images was 5,262 images, which was divided into a training set of 3,719 and a testing set of 1,543. The distribution numbers of each mulberry cultivar are shown in Fig. 2.

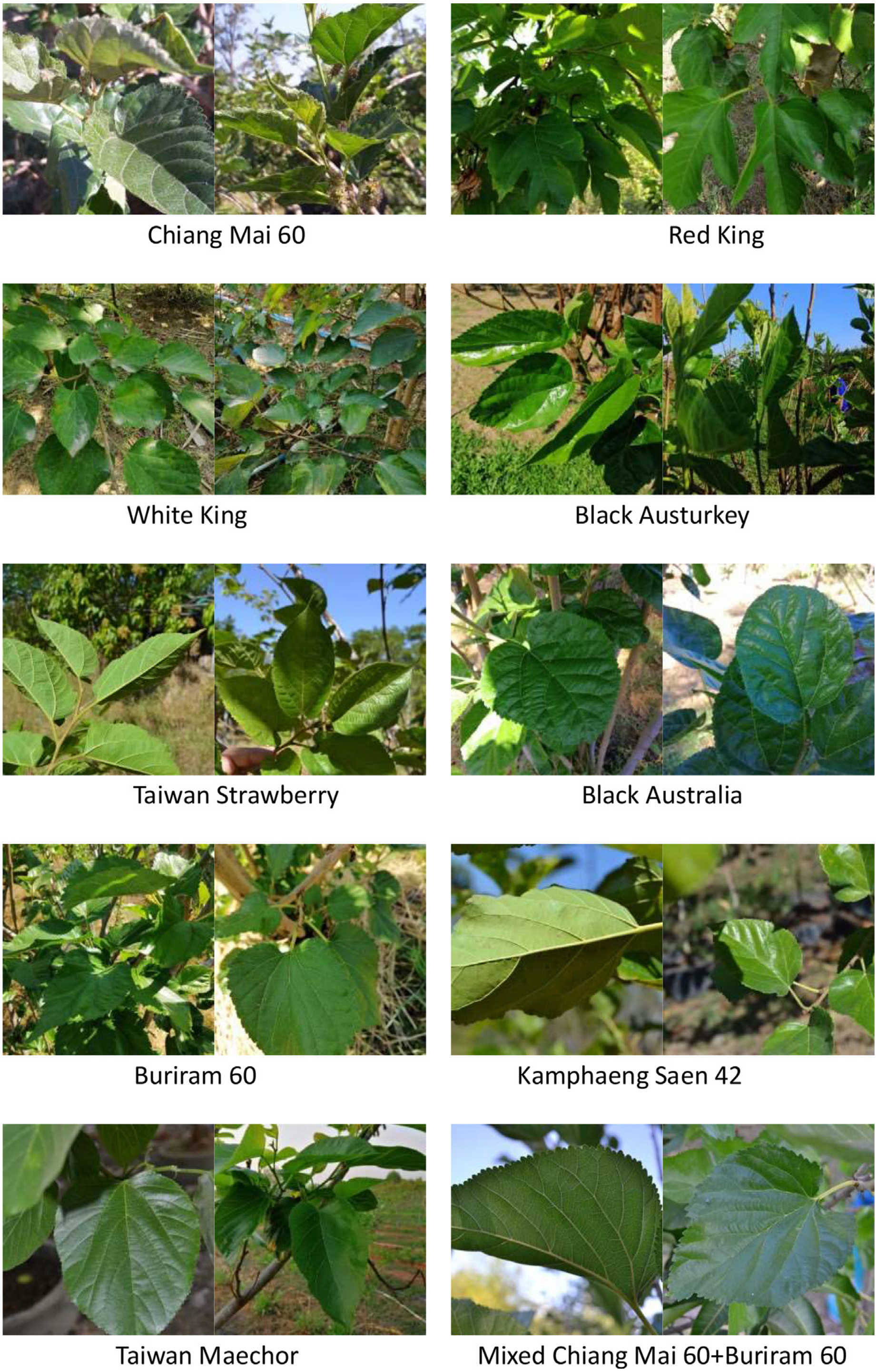
We provided more insights about the nature and possible variations of similar leaves with varying appearances arising due to the angular positioning of the camera when acquiring the mulberry leaves and the illumination variations due to contrast adjustments. This implies that the mulberry leaf dataset could contain a few similar leaves with variations in the background information, spatial representation, or illumination conditions in both training and test sets, as shown in Fig. 3. Moreover, many of the images within the dataset are dissimilar.

Note that, the mulberry leaf dataset reported in article [9] is available on Mendeley Data [11] and includes both a metadata file (Metadata – Mulberry Leaf Dataset.csv) and raw images (Mulberry Leaf Dataset directory), as shown in Fig. 4(a). The root directory of the mulberry leaf dataset contains two subdirectories (training and test set folders), as shown in Fig. 4(b). Each training set or test set contains ten subdirectories (classes), as shown in Fig. 5.

We uploaded the metadata file (Metadata – Mulberry Leaf Dataset.csv), which provides a structured tabular description of the collected dataset, as shown in Table 1. This table contains the following field columns: folder, class, filename, image type, pixel resolution (pixels), image size (KB), and color space (channels).

Table 1
Metadata description of the mulberry leaf dataset.

Folder	Class	Filename	Image type	Pixels	Size (KB)	Channels
Training set	01 ChiangMai60	2019207_105941.jpg	JPEG	(224, 224)	21	RGB
	01 ChiangMai60	2019207_105944.jpg	JPEG	(224, 224)	21	RGB
	01 ChiangMai60	2019207_105951.jpg	JPEG	(224, 224)	23	RGB
	01 ChiangMai60	2019207_110008.jpg	JPEG	(224, 224)	23	RGB
	01 ChiangMai60	2019207_110014.jpg	JPEG	(224, 224)	22	RGB
	01 ChiangMai60	2019207_110021.jpg	JPEG	(224, 224)	26	RGB
	01 ChiangMai60	2019207_1100238.jpg	JPEG	(224, 224)	25	RGB



Chiang Mai 60

Red King

White King

Black Austurkey

Taiwan Strawberry

Black Australia

Buriram 60

Kamphaeng Saen 42

Taiwan Maechor

Mixed Chiang Mai 60+Buriram 60

Fig. 1. Example images of ten mulberry leaf cultivars.

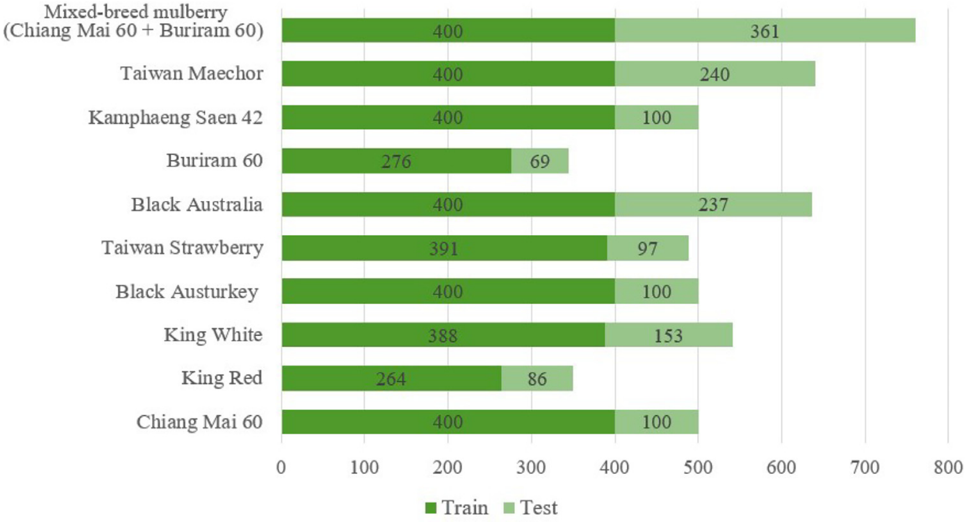


Fig. 2. The distribution numbers of training and test sets of ten mulberry cultivars in the mulberry leaf dataset.

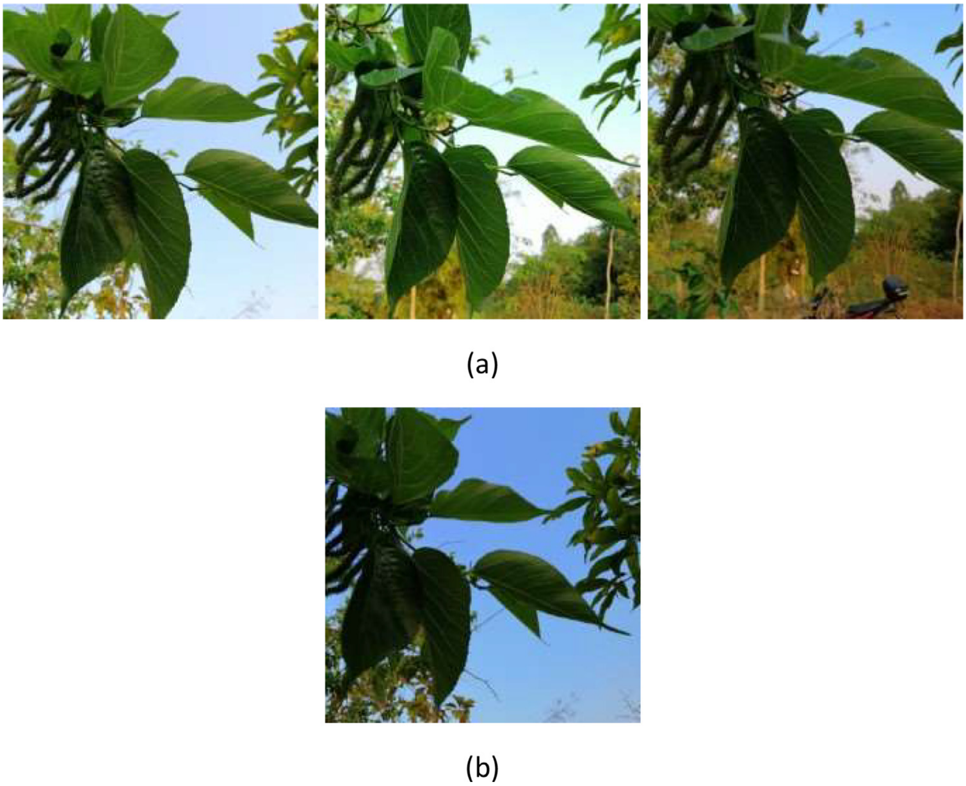
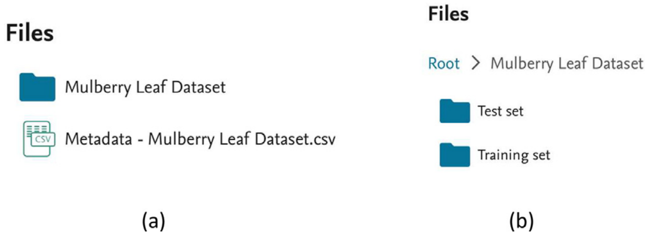


Fig. 3. An illustration of the same TaiwanMeacho leaves that can appear in both (a) training having different background conditions, illumination variations, and variation in angular view perspective and (b) test sets depicting variation in illumination and angular perspective views.



Files

Mulberry Leaf Dataset

Metadata - Mulberry Leaf Dataset.csv

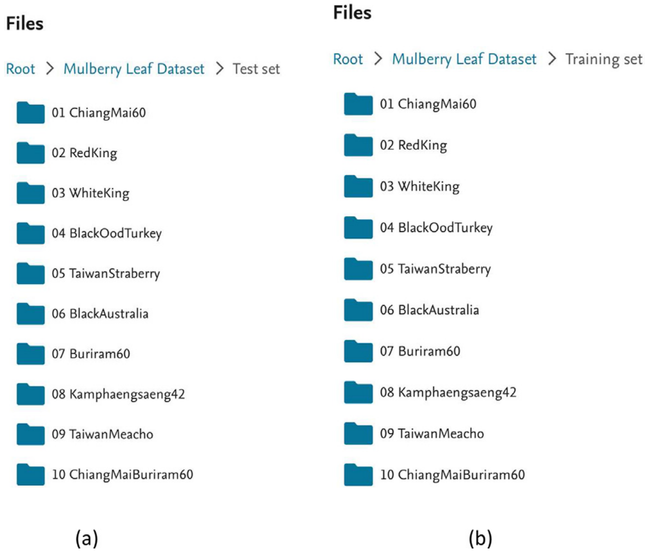
Files

Root > Mulberry Leaf Dataset

Test set

Training set

Fig. 4. An illustration of the data repository description of the mulberry leaf dataset on the Mendeley Data website: (a) Root directory and (b) directory of mulberry leaf dataset containing both test and training sets with its associated subdirectories (10 classes).



Files

Root > Mulberry Leaf Dataset > Test set

01 ChiangMai60
02 RedKing
03 WhiteKing
04 BlackOodTurkey
05 TaiwanStraberry
06 BlackAustralia
07 Buriram60
08 Kamphaengsaeng42
09 TaiwanMeacho
10 ChiangMaiBuriram60

Files

Root > Mulberry Leaf Dataset > Training set

01 ChiangMai60
02 RedKing
03 WhiteKing
04 BlackOodTurkey
05 TaiwanStraberry
06 BlackAustralia
07 Buriram60
08 Kamphaengsaeng42
09 TaiwanMeacho
10 ChiangMaiBuriram60

Fig. 5. A depiction of the test set (a) and training set (b) containing ten subdirectories (data classes).

4. Experimental Design, Materials and Methods

1. Data collection

We collected the mulberry leaf dataset from three main demographic regions in Thailand: Northern, Central, and Northeast. We presented our dataset containing the mulberry leaf images accessible via the Mendeley data repository (https://data.mendeley.com/datasets/ds45yy9jrc/3) [11].

2. Data acquisition

We curated the dataset between 2019 and 2020 from three main regions of Thailand: Northern, Central, and Northeast (within five provinces). In December 2019, we visited three provinces (Phitsanulok, Chiang Mai, and Nakhon Ratchasima) to capture images of mulberry leaves. Additionally, we acquired samples of the mulberry leaf images from Mahasarakham province between January and March 2020. In March 2020, we travelled to Buriram province to collect more images of mulberry leaves. Furthermore, we described the condition of the captured imagery of the mulberry leaves in the data description section.

The image captures of mulberry leaves were taken from five Thai areas, as shown in Fig. 6.

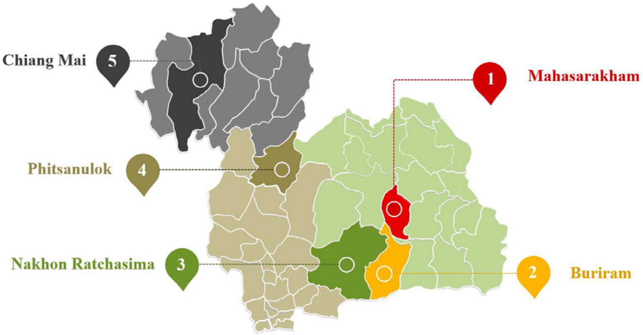


Fig. 6. The areas of mulberry fields in Thailand are described in five main provinces: Chiang Mai (Northern region), Phitsanulok (Center region), and Mahasarakham, Nakhon Ratchasima, and Buriram (Northeast region).

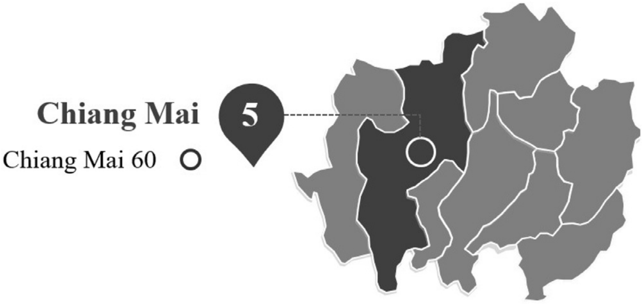


Fig. 7. Illustration of the mulberry cultivar collected in Chiang Mai province, Northern Thailand.

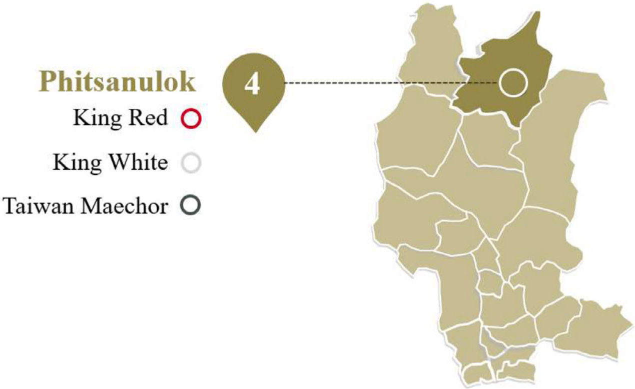


Fig. 8. Illustration of the mulberry cultivars collected in Phitsanulok province, the Center area of Thailand.

In Fig. 7, we gathered the Chiang Mai 60 cultivar from Chiang Mai province. The King Red, King White, and Taiwan Maechor cultivars were collected from the Phitsanulok (see Fig. 8). Additionally, we collected the Black Austurkey and Chiang Mai 60 cultivars from Nakhon Ratchasima. The King White and Black Austurkey cultivars were collected from Buriram. In Mahasarakham province, we also gathered ten mulberry cultivars, as shown in Fig. 9.

3. Data Cleaning

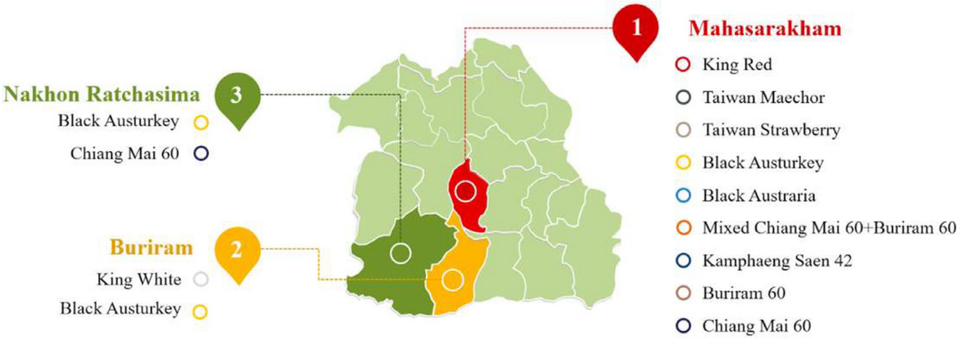


Fig. 9. Illustration of the provinces and mulberry cultivars collected in Northeast Thailand, consisting of Mahasarakham, Nakhon Ratchasima, and Buriram.

We inspected the curated imagery of the mulberry leaves. We ensured that all images containing blurred, noisy, or out-of-focus appearances were discarded from the dataset while maintaining imagery with high-quality resolution. Additionally, to foster uniformity in the image samples, we wrote a simple Python script that employs the scikit-image library package for resizing all the desired images within our dataset. The specified image size is 224 × 224 pixels. The described images within this dataset are accessible via Mendeley Data [11]. The computer vision researchers can use our proposed dataset for image classification or mulberry recognition tasks.

Limitations

Not applicable.

Ethics Statement

The authors have read and follow the ethical requirements for publication in Data in Brief and confirming that the current work does not involve human subjects, animal experiments, or any data collected from social media platforms.

Data Availability

Mulberry leaf dataset (Original data) (Mendeley Data).

CRediT Author Statement

Thipwimon Choompookham: Conceptualization, Data curation, Investigation, Methodology, Resources, Validation, Writing – original draft; **Emmanuel Okafor:** Conceptualization, Validation, Writing – review & editing; **Olarik Surinta:** Supervision, Conceptualization, Writing – review & editing.

Acknowledgements

This research project was financially supported by Mahasarakham University, Thailand.

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 paper.

References

[1] T. Munisami, M. Ramsurn, S. Kishnah, S. Pudaruth, Plant leaf recognition using shape features and colour histogram with k-nearest neighbour classifiers, Procedia Comput. Sci. 58 (2015) 740–747, doi:10.1016/j.procs.2015.08.095.

[2] N. Kumar, P.N. Belhumeur, A. Biswas, D.W. Jacobs, W.J. Kress, I.C. Lopez, J.V. Soares, Leafsnap: a computer vision system for automatic plant species identification, in: *European Conference on Computer Visio*n (ECCV 2012), 2012, pp. 502–516, doi:10.1007/978-3-642-33709-3_36.

[3] A. Ghosh, P. Roy, An automated model for leaf image-based plant recognition: an optimal feature-based machine learning approach, Innov. Syst. Softw. Eng. (2022), doi:10.1007/s11334-022-00440-y.

[4] S.S. Chouhan, U.P. Singh, A. Kaul, S. Jain, A data repository of leaf images: practice towards plant conservation with plant pathology, in: *4th International Conference on Information Systems and Computer Networ*ks (ISCON 2019), 2019, pp. 700–707, doi:10.1109/ISCON47742.2019.9036158.

[5] Hughes. D. P. & Salathe, M. (2015). An open access repository of images on plant health to enable the development of mobile disease diagonostics, *arXiv*, 1511.08060. DOI: 10.48550/arXiv.1511.08060

[6] S.S. Chouhan, A. Kaul, U.P. Singh, S. Jain, A database of leaf images: practice towards plant conservation with plant pathology, Mendeley Data (2019) V5, doi:10.17632/hb74ynkjcn.5.

[7] J. Hang, D. Zhang, P. Chen, J. Zhang, B. Wang, Classification of plant leaf diseases based on improved convolutional neural network, Sensors 19 (2019) 4161, doi:10.3390/s19194161.

[8] G. Geetharamani, J. Arun Pandian, Identification of plant leaf diseases using a nine-layer deep convolutional neural network, Comput. Electr. Eng. 76 (2019) 323–338, doi:10.1016/j.compeleceng.2019.04.011.

[9] T. Chompookham, O. Surinta, Ensemble methods with deep convolutional neural networks for plant leaf recognition, ICIC Express Lett. 15 (6) (2021) 553–565, doi:10.24507/icicel.15.06.553.

[10] P. Enkvetchakul, O. Surinta, Effective data augmentation and training techniques for improving deep learning in plant leaf disease recognition, Appl. Sci. Eng. Progr. 15 (3) (2022) 3810, doi:10.14416/j.asep.2021.01.003.

[11] T. Chompookham, O. Surinta, Mulberry leaf dataset, Mendeley Data V3 (2024), doi:10.17632/ds45yy9jrc.3.