ARTICLE IN PRESS
JID: DIB [mUS1Ga;October 6, 2024;1:16]
Data in Brief xxx (xxxx) xxx

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

A comprehensive dataset of rice field weed detection from Bangladesh

Q1 Q2 Md Sawkat Ali[a], Mohammad Rifat Ahmmad Rashid[a], Tasnim Hossain[b], Md Ahsan Kabir[b,*], Md. Kamrul[a], Sayam Hossain Bhuiyan Aumy[a], Mehedi Hasan Mridha[a], Imam Hossain Sajeeb[a], Mohammad Manzurul Islam[a], Taskeed Jabid[a]

[a] *Department of Computer Science and Engineering, East West University, Dhaka, Bangladesh*
[b] *Department of Electrical, Electronic and Communication Engineering, Military Institute of Science and Technology, Dhaka, Bangladesh*

ARTICLE INFO

Article history:
Received 6 June 2024
Revised 5 September 2024
Accepted 24 September 2024
Available online xxx

Dataset link: A comprehensive dataset of rice field weed detection from Bangladesh (Original data)

Keywords:
Automated weed detection
Rice field
Machine learning
Precision agriculture
Identification
Deep learning
Crop management

ABSTRACT

In agricultural research, particularly concerning rice cultivation, the presence of weeds within rice fields is acknowledged as a significant contributor to both diminished crop quality and increased production costs. Rice fields, due to their inherently moist environment, offer ideal conditions for weed proliferation. Traditionally, the control of these weeds has been managed through labor-intensive manual methods. However, as the agricultural sector evolves, there is a notable pivot towards leveraging advanced technological solutions, including deep learning and machine learning. The efficacy of these technologies hinges on the availability of high-quality, relevant data. To address this, a comprehensive dataset comprising 3632 high-resolution RGB images has been developed. This dataset is designed to capture a diverse range of weed species, specifically 11 types that are frequently found in rice fields. The diversity of the dataset ensures that machine learning models trained using this data can effectively identify and differentiate between desired and undesired plant species. While the dataset predominantly includes images from Bangladesh, the weed species it documents are

* Corresponding author.
E-mail address: ahsan@eece.mist.ac.bd (M.A. Kabir).

https://doi.org/10.1016/j.dib.2024.110981

Please cite this article as: M.S. Ali, M.R.A. Rashid and T. Hossain et al., A comprehensive dataset of rice field weed detection from Bangladesh, Data in Brief, https://doi.org/10.1016/j.dib.2024.110981

commonly found across various global rice-growing regions, enhancing the dataset's applicability in different agricultural settings.

Specifications Table

Subject	Precision Agriculture
Specific subject area	Computer vision techniques for the rice field weed detection and classification.
Type of data	Image Raw, and analyzed
Data collection	The original images were collected from fifteen distinct rice fields using high-resolution cameras to ensure quality and detail. To facilitate rapid learning by the system, a portion of these images was captured against a white background, providing a stark contrast that aids in quicker and more accurate identification. Every image was taken with precision, ensuring that extraneous objects were excluded, thereby maintaining the focus on the subject of interest. To encapsulate the full spectrum of variations that each object might exhibit, the images were taken with deliberate variations. This included manually adjusting the zoom to capture finer details, taking shots under different lighting conditions such as direct sunlight and the dappled shade of trees, and photographing the subjects from various angles.
Data source location	The following two locations are selected for data collection. Narsingdi, Dhaka, Bangladesh (Latitude: 23.920717° N, Longitude: 90.718811° E) Munshiganj, Dhaka, Bangladesh (Latitude: 23.5422° N, Longitude: 90.5305° E)
Data accessibility	Repository name: Mendeley data Data identification number: 10.17632/mt72bmxz73.4 Direct URL to data: https://data.mendeley.com/datasets/mt72bmxz73/4

1. Value of the Data

- Deep learning models trained on this comprehensive dataset can accurately differentiate between crop plants and various weed species, allowing for precise, targeted weed control that integrates seamlessly with automated systems for spot spraying of herbicides [2,3].
- Deep learning algorithms can analyze the dataset to uncover subtle patterns and insights that might be imperceptible to human observers, informing better farm management practices, predicting weed proliferation trends, and guiding the development of resistant crop strains [6].
- Once trained, these machine learning models can be deployed across various farms and regions, offering a scalable solution that adapts to different environmental and agricultural conditions in Bangladesh [7].
- By automating weed detection, these models drastically cut down the labor and time typically needed for weed management, thereby decreasing overall farm management costs [4,5].
- The dataset serves as a benchmark for the research community, not only within Bangladesh but globally. It encourages innovation in agricultural technology and provides a foundation for interdisciplinary research, combining agronomy, computer science, and environmental studies to tackle complex agricultural challenges [8,9].

2. Background

Bangladesh, spanning only 147,570 square kilometers with a population exceeding 166 million, relies heavily on agriculture, which has historically underpinned its achievement of food security despite frequent natural disasters and population growth. Projections suggest that by

2050, the agriculture-based economy could generate a gross domestic product of USD 3367 million [1]. However, as one of the most climate-vulnerable nations globally, Bangladesh faces significant threats to agriculture, particularly from flooding, salt intrusion, and drought. Rice cultivation, occupying about 75% of the total cropland and over 80% of all irrigated land [2], faces numerous challenges, including competition from weeds which reduce yields and degrade crop quality. In response, machine learning has emerged as a promising technology for developing automated weed management systems that can identify and control weeds in real-time, potentially enhancing production while reducing herbicide use. Despite various studies on crop diseases, there is a lack of datasets specifically for rice weeds in Bangladesh. This paper presents a novel dataset focused on rice weeds, offering detailed insights into various weed types and their leaf traits, which could significantly enhance the precision of machine learning applications in agriculture.

3. Data Description

The rice field weeds dataset is a collection of standard images, comprising 3632 pictures of 11 types of weeds found in various rice fields in Bangladesh. These 11 weeds were identified through field research and consultation with local agricultural officials. After that, 11 different types were titled after the scientific name of the weeds. Table 1 shows the types of rice weeds those image data are collected and Fig.1 presents the image of eleven different types of collected data. The uniqueness of this dataset lies in the fact that it is the first of its kind, with no other collection of images related to rice field weeds previously recorded.

Paspalum scrobiculatum [12], known as kodo millet or kodo grass, is often considered a weed in rice fields. It is part of the Paspalum genus, which includes many species that are considered troublesome in agricultural settings. However, it is also cultivated for its grain in various

Table 1
Types of weeds of rice.

1	Paspalum scrobiculatum	7	Alternanthera philoxeroides
2	Ipomoea aquatic	8	Synedrella nodiflora
3	Cyperus ochraceus	9	Fimbristylis littoralis
4	Centella asiatica	10	Commelina benghalensis
5	Panicum repens	11	Pteris vi ttata
6	Marsilea minuta		

Fig. 1. Paspalum scrobiculatum.

Please cite this article as: M.S. Ali, M.R.A. Rashid and T. Hossain et al., A comprehensive dataset of rice field weed detection from Bangladesh, Data in Brief, https://doi.org/10.1016/j.dib.2024.110981

Fig. 2. Ipomoea aquatic.

Fig. 3. Cyperus ochraceus.

parts of the world, especially in India, where it's valued for its drought resistance and nutritional benefits. Fig. 1 illustrates a sample image of Paspalum scrobiculatum from bangladesh rice fields.

Ipomoea aquatica [12], known as water spinach, swamp cabbage, or kangkong, is an aquatic or semi-aquatic plant widely grown in Southeast Asia and some parts of the Southern United States. It thrives in moist and aquatic environments, making it a popular vegetable crop in regions with waterlogged conditions. Fig. 2 illustrates a sample image of Paspalum scrobiculatum from bangladesh rice fields.

Cyperus ochraceus [12] is a species of sedge found in wetlands and moist habitats. It belongs to the Cyperaceae family, which includes many species commonly referred to as sedges. The plant is notable for its ability to thrive in both semi-aquatic environments and damp terrestrial areas (Figs. 3–11).

Centella asiatica [12], known as Gotu Kola, is a perennial herbaceous plant that thrives in and around water. It is native to the wetlands of Asia, widely used in the cuisines and traditional medicines of several cultures, particularly within South and Southeast Asia.

Panicum repens [12], known as torpedograss, is a perennial grass species native to Africa and Eurasia but has become widely distributed in tropical and subtropical regions around the world, often categorized as an invasive weed. It thrives in wetlands, along water bodies, and

Please cite this article as: M.S. Ali, M.R.A. Rashid and T. Hossain et al., A comprehensive dataset of rice field weed detection from Bangladesh, Data in Brief, https://doi.org/10.1016/j.dib.2024.110981

Fig. 4. Centella asiatica.

Fig. 5. Panicum repens.

in irrigated agricultural fields, posing challenges due to its aggressive growth and robust root system.

Marsilea minuta[12], known as dwarf waterclover, is a species of aquatic fern in the family Marsileaceae. It is found in various parts of the world, including parts of Asia, Africa, Australia, and the Americas. This plant is noted for its small size and its clover-like leaves, which makes it distinct in appearance and a popular choice for aquariums and water gardens.

Alternanthera philoxeroides[12], commonly known as alligator weed, is a perennial aquatic plant that originates from South America but has spread to various parts of the world, often considered an invasive species outside its native range. It thrives in aquatic and semi-aquatic environments such as lakes, rivers, and marshes, as well as on moist land.

Synedrella nodiflora[12], known as nodeweed or Cinderella weed, is a tropical and subtropical flowering plant belonging to the family Asteraceae. It is widespread across various regions, including parts of Africa, Asia, and the Americas, where it typically grows in disturbed areas, road sides, and cultivated fields.

Fimbristylis littoralis [12], known as the lesser fimbry or fringe-rush, is a species of sedge belonging to the family Cyperaceae. It is found in a variety of wet habitats across tropical and subtropical regions worldwide, including marshes, wetlands, and the edges of ponds and streams.

Commelina benghalensis [12], known as Benghal dayflower, tropical spiderwort, or wandering Jew, is a perennial herb that is native to tropical Asia and Africa but has become widespread

Please cite this article as: M.S. Ali, M.R.A. Rashid and T. Hossain et al., A comprehensive dataset of rice field weed detection from Bangladesh, Data in Brief, https://doi.org/10.1016/j.dib.2024.110981

Fig. 6. Marsilea minuta.

Fig. 7. Alternanthera philoxeroides.

Fig. 8. Synedrella nodiflora.

Please cite this article as: M.S. Ali, M.R.A. Rashid and T. Hossain et al., A comprehensive dataset of rice field weed detection from Bangladesh, Data in Brief, https://doi.org/10.1016/j.dib.2024.110981

Fig. 9. Fimbristylis littoralis.

Fig. 10. Commelina benghalensis.

in other tropical and subtropical regions as an invasive species. It is particularly prevalent in agricultural fields where it can be a troublesome weed.

Commelina benghalensis[12], known as Benghal dayflower, tropical spiderwort, or wandering Jew, is a perennial herb that is native to tropical Asia and Africa but has become widespread in other tropical and subtropical regions as an invasive species. It is particularly prevalent in agricultural fields where it can be a troublesome weed.

Table 2 provides a comparative overview of various agricultural datasets from the literature alongside our novel dataset on rice field weeds in Bangladesh. Datasets by Sudars et al. [6] and Krestenitis et al. [7] provide valuable insights into weed and crop management but are more generalist, covering multiple crop types and less focused on specific regional challenges. The size of our dataset is competitive, offering a large number of images similar to Ahmed et al. [8] and other studies, but with a concentrated focus on a single crop type, enhancing its utility for specific agricultural applications.

Fig. 11. Pteris vittata.

Table 2
Comparative overview of agricultural datasets highlighting crop and weed types, and dataset sizes from various studies.

Reference	Type of classes	Size of the dataset
[6]	6 crops, 8 weed species	1118 annotated images
[7]	Weeds in cotton fields	201 RGB images
[8]	Mango leaf health and disease	4000 digital RGB images
[9]	Soybean health, pest damage	6410 images
[10]	Citrus fruit and leaf health/disease	759 digital images
[11]	Maize, beans, leeks (early growth)	2801 images
Our Dataset	**Rice field weeds**	**3632 images of 11 weed types**

Table 3
Rice Field Weeds dataset information at a glance.

Type of data	RGB digital images
Data format	.jpg
Number of images	3632
Number of classes	11
How data were acquired	Through smartphone camera
Data source location	Rice fields of Narsingdi and Munshiganj, Bangladesh
Where applicable	Deep learning, image processing, agricultural studies

In November and December 2022, our research team conducted an extensive data collection initiative on rice field weeds in Bangladesh, identifying eleven major weed species and amassing over 4300 samples. The data collection employed multiple image-capturing techniques to enhance the dataset's utility. Initially, raw images were captured under various lighting conditions, including direct sunlight and shaded environments, to ensure consistency across different natural settings. Additionally, we took images against a controlled white background to eliminate extraneous visual elements, focusing exclusively on the weeds. Table 3 provides essential details about the dataset, including the type of data, format, total number of images, number of classes, methods of data acquisition, and the specific locations of data collection. Table 4 lists each weed class, the scientific names of the weeds, the number of images collected for each class, and the corresponding folder names where these images are stored. It serves as a detailed directory for navigating the dataset.

Please cite this article as: M.S. Ali, M.R.A. Rashid and T. Hossain et al., A comprehensive dataset of rice field weed detection from Bangladesh, Data in Brief, https://doi.org/10.1016/j.dib.2024.110981

Table 4
Folder distribution.

Weed class	Name of the disease	No of images	Folder Name
1	Alternanthera philoxeroide	106	W_CL_01_Alternanthera philoxeroide
2	Centella asiatica	404	W_CL_02_Centella asiatica
3	Commelina benghalensis	316	W_CL_03_Commelina benghalensis
4	Cyperus ochraceus	190	W_CL_04_Cyperus ochraceus
5	Fimbristylis littoralis	168	W_CL_05_Fimbristylis littoralis
6	Ipomoea aquatic	897	W_CL_06_Ipomoea aquatic
7	Marsilea minuta	346	W_CL_07_Marsilea minuta
8	Panicum repens	294	W_CL_08_Panicum repens
9	Paspalum scrobiculatum	345	W_CL_09_Paspalum scrobiculatum
10	Pteris vittata	221	W_CL_10_Pteris vittata
11	Synedrella nodiflora	345	W_CL_11_Synedrella nodiflora
	Total	3632	

4. Experimental Design, Materials and Methods

This research work focuses on the collection of 11 major types of rice weed images from different regions of Bangladesh considering the necessary parameters like soil, weather etc. Samples in this dataset are carefully curated by their textural descriptors and color scheme of different types of rice field weeds. To improve quality the images are then fine-tuned using different filters for noise reduction after the collection of raw images. Necessary steps were taken to ensure image variation that makes this one more accurate in terms of detection. Fig.12 presents the methodological flow diagram of the study.

4.1. Initial field assessment

In this study, field assessments are carried out to get knowledge on rice fields and weeds. In addition, advice is taken from local agricultural officials to identify the different weed species present in rice fields, their appearance, and the specific conditions under which each species thrives. After that, this information is used to develop targeted weed management strategies that are customized to the unique characteristics of each weed species. Throughout the acquisition of this knowledge, a notable observation emerged regarding substantial variation within weed species across diverse fields. This underscores the criticality of adopting a meticulous and methodical approach to weed management, recognizing that distinct weed populations may necessitate tailored strategies for effective control.

4.2. Rice fields selection

To ensure a comprehensive and diverse dataset, high-quality data from various regions of the country is important. Thus, a strategic approach was taken to select the rice fields for data collection. Various factors, including soil condition, land position, and sunlight disparity, have been analyzed when selecting over 15 rice fields across the Narsingdi and Munshiganj, two main districts of Bangladesh for agriculture production. These areas are privately owned and possess a flat terrain, making them ideal for farming activities. At various times of the year different varieties of rice grow in these fields, and local farmers cultivate it extensively. It provides sufficient opportunities to capture images of various stages of rice growth and weed infestation. The images were captured during the dry season when the water supply in the fields was low. As weeds require adequate water for growth, the availability of weeds during this time was limited. This created a challenge for collecting images of all weed species in equal numbers, as certain weeds may have been more prevalent during other seasons or when the water supply was

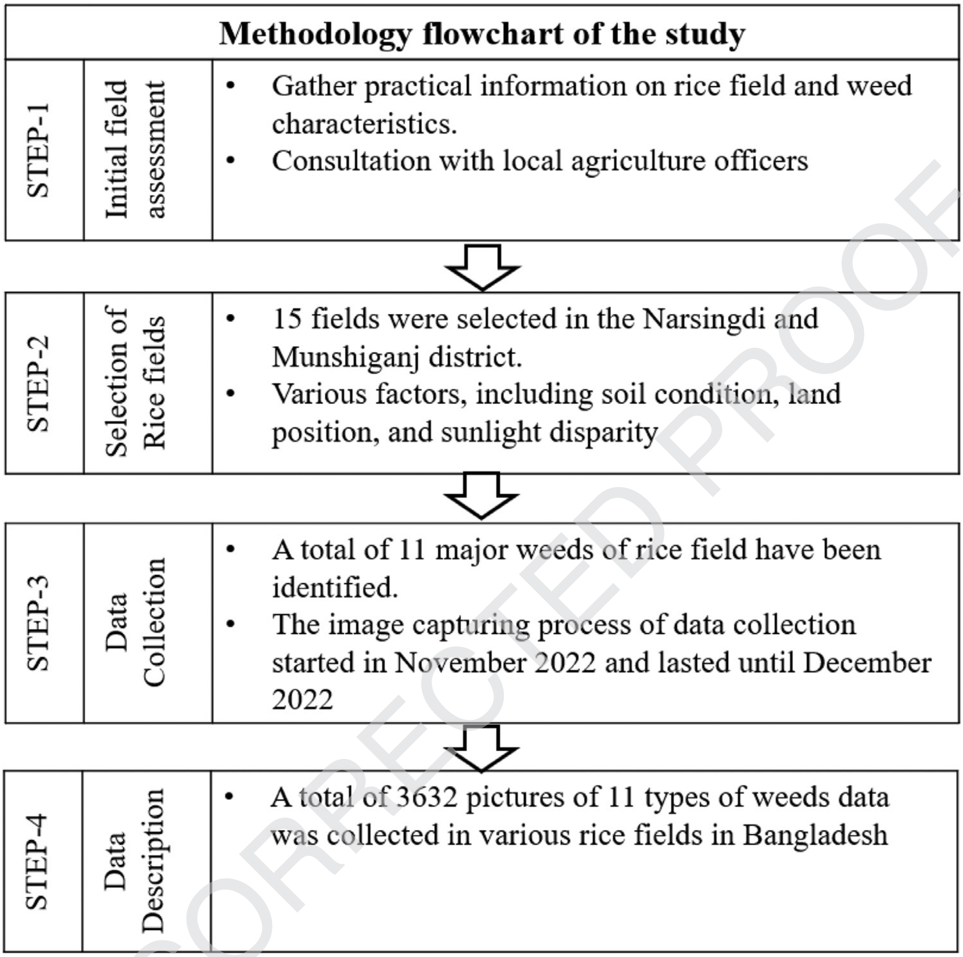

Fig. 12. Methodological flow diagram of the study.

higher. Despite these challenges, an extraordinary effort was given to capture a diverse range of weed species by carefully selecting the fields and capturing images from various angles and lighting conditions.

4.3. Data collection

During this study, smartphones equipped with a variety of megapixel lenses—including 64MP, 50MP, 13MP, 12MP, 8MP, and 2MP—were utilized to photograph several weed species. To capture diverse variations of each species, images were taken from multiple viewpoints and under different lighting conditions. This included manually adjusting the zoom to capture finer details and photographing under both direct sunlight and tree shade, as well as against both raw and controlled white backgrounds. The inclusion of a white background created a consistent environment that facilitated the system's ability to learn the distinct features of each weed species more effectively. Conversely, raw images were captured directly in the rice fields. Additional meticulous techniques, such as varying the camera angles and zoom levels, were employed to ensure comprehensive documentation of each species' subtle differences. These efforts were crucial for capturing the nuanced variations among the weeds, thereby enriching the dataset's utility for machine learning applications.

Limitations

The dataset of rice field weeds has several limitations, largely influenced by environmental factors affecting the growth of various weed species. For instance, Cyperus ochraceus and Centella asiatica are predominantly seen during the dry season in Bangladesh, whereas Alternanthera philoxeroides and Fimbristylis littoralis thrive in more humid conditions. Since all 3632 images in our dataset were collected during the dry season, the prevalence of certain weeds was naturally lower due to restricted water availability. This resulted in an uneven representation of weed types, with some species absent due to their scarce occurrence in the dry season. Despite these limitations, we consider them to be relatively minor and addressable in future datasets through more balanced collection efforts across different seasons.

Efforts to prepare a well-rounded dataset demand a diverse range of images, which are essential for machine learning research. This process requires considerable effort, dedication, expertise, and meticulous attention to detail. The main challenges encountered during dataset preparation included:

- The cultivation of rice in waterlogged fields leads to a proliferation of weeds, prompting us to identify 11 species that demonstrate significant growth under such conditions.
- Due to the timing of the dry season collection period in November and early December, it was challenging to gather a comparable number of images for each weed category.
- Fieldwork in rice paddies necessitated caution due to the risk of encountering poisonous insects and snakes.
- Adverse environmental conditions often complicated the process of capturing high-quality images.

Ethics Statement

The data is not gathered using the internet or social media.

Data Availability

A comprehensive dataset of rice field weed detection from Bangladesh (Original data) (Mendeley Data)

CRediT Author Statement

Md Sawkat Ali: Conceptualization, Methodology, Investigation, Data curation, Visualization, Writing – original draft; **Mohammad Rifat Ahmmad Rashid:** Conceptualization, Methodology, Supervision, Project administration, Validation, Writing – review & editing; **Tasnim Hossain:** Conceptualization, Methodology, Visualization, Writing – original draft, Writing – review & editing; **Md Ahsan Kabir:** Investigation, Data curation, Writing – original draft, Writing – review & editing, Validation; **Md. Kamrul:** Investigation, Data curation; **Sayam Hossain Bhuiyan Aumy:** Investigation, Data curation; **Mehedi Hasan Mridha:** Data curation, Writing – review & editing, Visualization; **Imam Hossain Sajeeb:** Writing – review & editing, Visualization; **Mohammad Manzurul Islam:** Conceptualization, Methodology, Supervision; **Taskeed Jabid:** Methodology, Writing – review & editing, Supervision.

Please cite this article as: M.S. Ali, M.R.A. Rashid and T. Hossain et al., A comprehensive dataset of rice field weed detection from Bangladesh, Data in Brief, https://doi.org/10.1016/j.dib.2024.110981

Acknowledgments

We extend our sincere gratitude to Kbd. Md. Abu Saeed Shubvro, Upazilla Agriculture Officer, for his invaluable expertise and guidance in the annotation of the dataset. We are profoundly grateful for his contributions, which were instrumental in enhancing the quality of our work. We would also like to thank our field assistants who participated diligently in the data collection and preliminary annotation phases.

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] Bangladesh Rice Research Institute, “Rice in Bangladesh,” 2023. http://www.knowledgebank-brri.org/riceinban.php (accessed 08 April 2023).
[2] B. Dey, M. Masum Ul Haque, R. Khatun, R. Ahmed, Comparative performance of four CNN-based deep learning variants in detecting Hispa pest, two fungal diseases, and NPK deficiency symptoms of rice (Oryza sativa), Comput. Electron Agric. 202 (2022) 107340.
[3] M.S.H. Talukder, A.K. Sarkar, Nutrients deficiency diagnosis of rice crop by weighted average ensemble learning, Smart Agric. Technol. 4 (2023), doi:10.1016/j.atech.2022.100155.
[4] M.A.B. Bhuiyan, H.M. Abdullah, S.E. Arman, S. Saminur Rahman, K. Al Mahmud, BananaSqueezeNet: a very fast, lightweight convolutional neural network for the diagnosis of three prominent banana leaf diseases, Smart Agric. Technol. 4 (2023), doi:10.1016/j.atech.2023.100214.
[5] K. Sudars, J. Jasko, I. Namatevs, L. Ozola, N. Badaukis, Dataset of annotated food crops and weed images for robotic computer vision control, Data Br. 31 (2020) 105833, doi:10.1016/j.dib.2020.105833.
[6] M. Krestenitis, et al., CoFly-WeedDB: a UAV image dataset for weed detection and species identification, Data Br. 45 (2022) 108575, doi:10.1016/j.dib.2022.108575.
[7] S.I. Ahmed, et al., MangoLeafBD: a comprehensive image dataset to classify diseased and healthy mango leaves, Data Br. 47 (2023) 108941, doi:10.1016/j.dib.2023.108941.
[8] M.E. Mignoni, A. Honorato, R. Kunst, R. Righi, A. Massuquetti, Soybean images dataset for caterpillar and Diabrotica speciosa pest detection and classification, Data Br. 40 (2022) 107756, doi:10.1016/j.dib.2021.107756.
[9] H.T. Rauf, B.A. Saleem, M.I.U. Lali, M.A. Khan, M. Sharif, S.A.C. Bukhari, A citrus fruits and leaves dataset for detection and classification of citrus diseases through machine learning, Data Br. 26 (2019) 104340, doi:10.1016/j.dib.2019.104340.
[10] L. Lac, B. Keresztes, M. Louargant, M. Donias, J.-P.Da Costa, An annotated image dataset of vegetable crops at an early stage of growth for proximal sensing applications, Data Br. 42 (2022) 108035, doi:10.1016/j.dib.2022.108035.
[11] Y. Suryawanshi, K. Patil, P. Chumchu, VegNet: dataset of vegetable quality images for machine learning applications, Data Br. 45 (2022) 108657, doi:10.1016/j.dib.2022.108657.
[12] B.P Caton, A practical field guide to weeds of rice in Asia, Int. Rice Res. Inst. (2010).