Data in Brief 54 (2024) 110334

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

Grain rot dataset caused by Burkholderia Glumae Bacteria

Khang Nguyen Quoc, Luyl-Da Quach*

FPT University, Can Tho campus, Cantho city, Vietnam

A R T I C L E I N F O

Article history:
Received 9 February 2024
Revised 11 March 2024
Accepted 12 March 2024
Available online 16 March 2024

Dataset link: Bacterial Grain Rot Dataset (Original data)

Keywords:
Bacterial grain
Object detection
Image processing
Computer vision
Rice disease

A B S T R A C T

The Burkholderia glumae bacterium causes bacterial grain rot in rice, posing significant threats to the crop's yield, particularly thriving during the rice flowering and grain filling stages. This disease is especially evident in rice grains before harvest, presenting challenges in the detection and classification of rice panicles. Firstly, diseased grains may mix with healthy ones, complicating their separation. Secondly, the size of grains on a panicle varies from small to large, which can be problematic when detected using object detection methods. Thirdly, disease classification can be conducted by evaluating the extent of infection on rice panicles to assess its impact on yield. Finally, the challenges in detection, classification, and preprocessing for disease identification and management necessitate the adoption of diverse approaches in machine learning and deep learning to develop optimal methods and support smart agriculture

* Corresponding author.
E-mail address: daql@fe.edu.vn (L.-D. Quach).

https://doi.org/10.1016/j.dib.2024.110334

Specifications Table

Subject	Agricultural Sciences
Specific subject area	Image Processing, Smart agriculture, Image Identification, object detection, Image classification, computer vision, artificial intelligence, and deep learning.
Data format	Raw Image
Type of data	Image
Data collection	The data was collected by the author in rice fields in the Mekong Delta region using a mobile phone.
Data source location	Provinces in the Mekong Delta Latitude: 10.063363, Longitude: 105.594339
Data accessibility	Repository name: Bacterial Grain Rot Dataset Caused by Burkholderia Glumae Bacteria Data identification number: 10.5281/zenodo.10805462 Direct URL to data: https://zenodo.org/records/10805462 Guidance on retrieving this dataset: Individuals may obtain the dataset by downloading it from the provided link and then unzipping the files for use.
Related research article	Quach, Luyl-Da, et al. "Evaluating the Effectiveness of YOLO Models in Different Sized Object Detection and Feature-Based Classification of Small Objects." Journal of Advances in Information Technology 14.5 (2023): 907–917. ⟨13:italic ⟩https://www.jait.us/show-232–1397–1.html⟨/13:italic⟩.

1. Value of the Data

- Overview: The dataset is approximately 341.8 MB in size and contains 1528 images across two object classes, along with labeled text files. Each class consists of various images of healthy and diseased rice grains, all consistently sized at 1280 × 1280 pixels. The data has been processed to enhance diversity in brightness, angles, and to clarify edge features, improving the model's recognition efficiency.
- First Open-Data Access: This dataset is the first publicly released dataset related to Bacterial Grain Rot based on object detection. It has accelerated advancements in disease detection, monitoring, and management in rice production through collaboration among researchers.
- Potential in New Approach for Bacterial Grain Rot Disease: Utilizing this data, several studies have been able to automate the diagnosis of disease severity in rice. Identifying and quantifying the proportion of diseased rice grains per unit can provide a clear measure of disease severity. Based on this method, a smart farming system can be constructed, allowing farmers to automatically monitor disease levels and implement the most effective management strategies, such as optimizing pesticide application.
- Precision Agriculture Applications: The dataset has been used in studies [1,2], yielding promising results in identifying diseased rice grains with bacterial grain rot on panicles. However, the unresolved issue in these studies is the application of YOLOv5, YOLOv6, and YOLOv7 models for evaluation, with the highest accuracy achieved around 90%, and classification based on the characteristics of bacterial grain rot achieving over 70%. This presents numerous challenges that need to be addressed to improve accuracy and object recognition.
- Anticipation is high that this dataset will lead to beneficial results as it undergoes additional scholarly investigation and expert assessment within the agricultural sector. This aspect is vital, considering that different nations assess the disease's severity (and its effect on crop yields) using distinct measurement criteria.

2. Background

The bacterial pathogen Burkholderia glumae is the cause behind grain rot in rice, which notably affects the crop's yield [3]. This bacterium thrives at different developmental phases of

the rice, leading to rot in seedlings, blight in leaves, and rot in grains, especially during the bloom phase. This condition is acknowledged as a critical agricultural disease in various nations spanning Asia, Central America, South America, and South Africa. The disease can reduce rice yields by up to 75%, as reported in certain regions of the United States. It is recognized as a critical disease in countries such as Japan [4], Thailand [5], Iran [6], and South Korea among others [7].

The disease can be identified by small, uneven brown spots on the flag leaf, and infected grains may appear from light to dark brown, rotten, shrivelled, and partially filled [8]. Disease diagnosis primarily relies on manual methods or laboratory techniques such as PCR [9], ELISA [10], and estimating infection risk using weather data. However, these methods are costly and challenging to apply in practical production settings. Consequently, some studies have explored computer vision techniques for disease detection, as demonstrated in research [1,2]. Additionally, new and diverse machine learning techniques have been developed and applied to detect diseases on rice leaves, as shown in studies [11–13]. This underscores the necessity for datasets that enable further research using various approaches.

Importance of Grain Rot

- Economic Impact: Grain rot directly affects the rice grains, which are the harvestable part of the plant. This symptom has a significant economic impact because it directly correlates with the marketable yield of the crop.
- Yield Loss: Among the symptoms, grain rot can lead to severe yield losses as it affects the development and viability of the rice grains. In severe infections, grain rot can lead to complete loss of the panicle, which is the primary yield component of rice.
- Quality of Produce: Grain rot not only reduces the quantity but also the quality of the rice grains. Infected grains are often discolored, unmarketable, and unsuitable for consumption, affecting the overall quality of the rice produced.

3. Data Description

In the agricultural sector, image datasets are of paramount importance, aiding in fields ranging from computer vision and machine learning to the development of intelligent agricultural systems. They offer a substantial repository of practical data that researchers utilize to confirm the efficacy of various models, algorithms, and theoretical frameworks. Moreover, the use of wide-ranging and rigorously reviewed image datasets affords researchers the opportunity to discover novel methods that enhance the results of prior studies or to experiment with innovative research techniques.

The causes of grain rot in rice can originate from spider mites, fungi, and particularly from the bacterium Pseudomonas glumae (also known as Burkholderia glumae). The characteristic symptom of rice infected by Burkholderia glumae is the blackening of the grains or the presence of disease spots on the grain husk. This feature facilitates the identification of the disease through image processing techniques or artificial intelligence more easily compared to other methods [14].

Bacterial grain rot disease in rice significantly affects crop yield. Such datasets offer researchers, agronomists, and farmers valuable resources for identifying, classifying, and studying rice leaves. Through data analysis, new models can help speed up disease detection and improve the accuracy of disease severity identification on rice leaves, thereby ensuring widespread rice productivity. In summary, the Grain Rot Dataset plays a vital role in advancing research, enhancing smart farming, and ensuring the overall health and productivity of rice crops.

With original data collected from the Mekong Delta region, we have processed the data. Ultimately, the Bacterial Grain Rot Dataset comprises 1528 high-resolution images uniformly sized at 1280 $\times$ 1280 pixels in JPEG format. The data have been labelled and bounding boxes assigned for all objects within two main classes: healthy grains and diseased grains. The data are divided into three folders representing the training, validation, and testing sets, with each set contain-

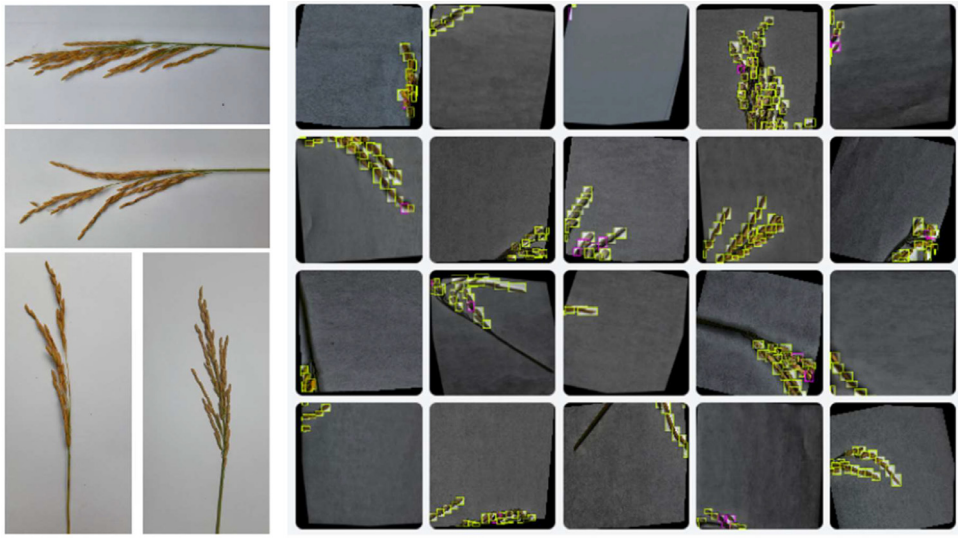

Fig. 1. Image data sample in dataset.

ing images and bounding box label files. Fig. 1 shows examples of original and processed rice images in the dataset.

4. Experimental Design, Materials and Methods

4.1. Field data collection

The data were collected by the authors from various rice fields in the provinces of the Mekong Delta region through sampling and evaluation methods. The collected samples were securely wrapped and then photographed using a mobile device. The method employed involved taking pictures on a white background, with a fixed distance maintained between the phone and the sample, as illustrated in Fig. 2. As a result, the collected data includes 566 original images, high-quality images with a resolution of up to 96 dpi. Additionally, the number of rice grains

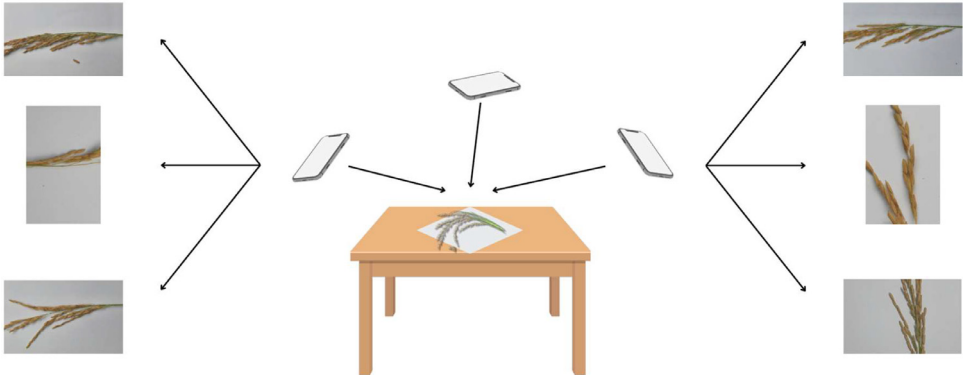

Fig. 2. Method of image acquisition.

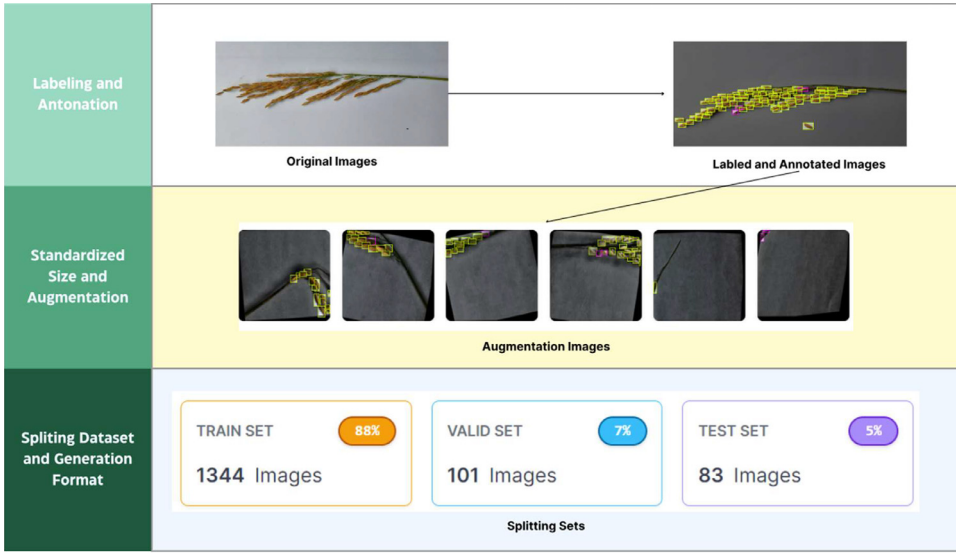

Fig. 3. Process of image prepocessing.

Table 1
Statistics on the number of healthy and diseased instance (seeds) processed.

Classes	Healthy	Disease	Total
Number of seed	11,287	1189	12,467

in each image varies widely, ranging from 18 to 161 grains per image, and the total number of objects summed up to more than 12,400 instances (seed), detailed in Fig. 3.

4.2. Data preprocessing

The process of handling the collected images was carried out in steps, including: (1) Labeling and Annotation, (2) Standardizing Size and Augmentation, and (3) Splitting the Dataset and Generating Format, as illustrated in Fig. 3.

Step 1: Labelling and Annotation: After collection, the data are uploaded to RoboFlow for labelling and assigning boxes to each rice grain in the images. Table 1 summarizes the number of objects labelled in this step. The distribution of total objects from each class per image is shown in Fig. 4.

Step 2: Standardized Size and Augmentation: During this step, labelled and annotated images are resized to the standard dimension required by recognition models, which is 1280 $\times$ 1280 pixels. Subsequently, many transformations are applied to increase the quantity of images and diversify their features, including Auto-Adjust Contrast and Brightness to balance differences between light and dark areas in the images, enhancing object recognition on white backgrounds; and Flipping, Rotating, and Shearing to generate more diverse image angles. After processing, the dataset includes over 1500 images with a variety of brightness levels, angles, and object counts.

Step 3: Splitting the Dataset and Generating Format: The final data are divided into three sets consisting of a training set, validation set, and testing set, to facilitate the training and testing of object recognition models. Table 2 displays the number of images within these three sets with a ratio of 8:1:1.

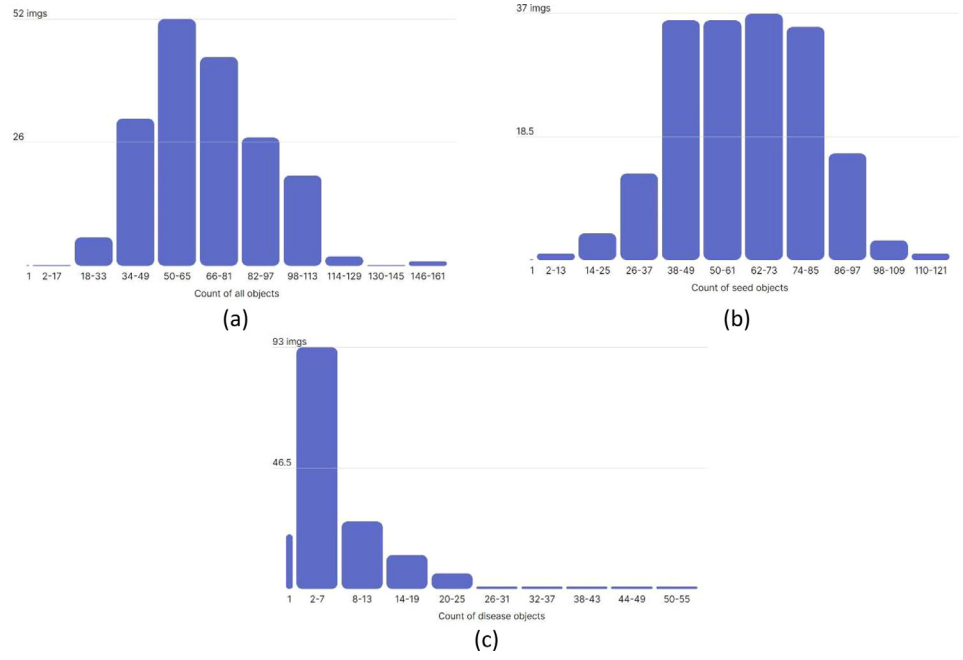

Fig. 4. Distribution of total objects: (a) Total objects, (b) Heathy Class and (c) Disease Class.

Table 2
Statistics on the number of grain rot disease images in the dataset.

Dataset	Training Set	Validation Set	Testing Set
Nunmber of images	1344	101	83

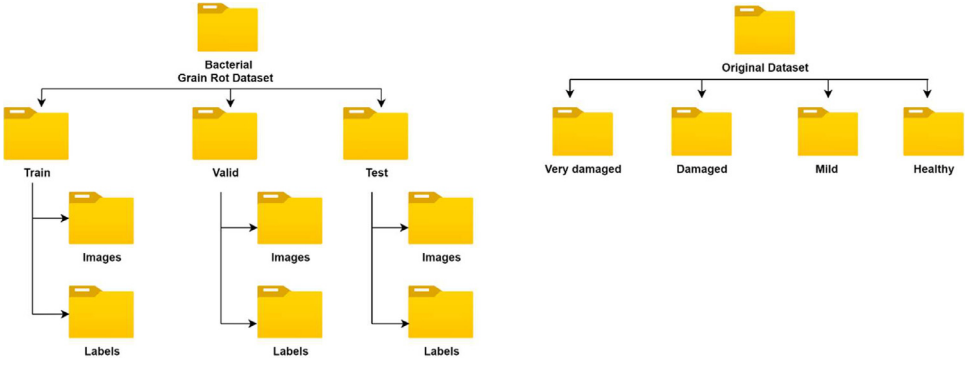

Fig. 5. Arrangement of the Bacterial Grain Rot Dataset dataset's folders.

This data is generated in the YOLO format, and for other studies, the YOLO model can be used by reformatting with RoboFlow or LabelImg. Besides, the original data set is organized into 4 disease levels according to the hierarchy of Vietnamese standards. The structure of the dataset model is illustrated in Fig. 5.

Limitations

Not applicable.

Ethics Statement

This study did not conduct experiments involving humans and animals.

CRediT Author Statement

Khang N. Quoc: Visualization, Writing - original draft. **Luyl-Da Quach:** guided the execution process, methodology, writing, and investigation of related information, and performed the final review and edits on the manuscript.

Data Availability

Bacterial Grain Rot Dataset (Original data) (10.5281/zenodo.10805462)

Acknowledgements

The authors would like to thank the support of farmers in the Mekong Delta who assisted us in data collection. Besides, we would also like to thank the support from expert Dr. Chau La Hoang (ATREM Limited Liability Company) for consulting and providing comments during the data collection process. Email of company representative: hoangchau@ctu.edu.vn

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] L.-D. Quach, K.N. Quoc, A.N. Quynh, H.T. Ngoc, Evaluating the effectiveness of YOLO models in different sized object detection and feature-based classification of small objects, J. Adv. Inf. Technol. 14 (5) (2023).
[2] K. Nguyen Quoc, A. Nguyen Quynh, H.Tran Ngoc, L.-D. Quach, Using the New YoLo Models in Detecting Small-Sized Objects in the Case of Rice Grains on Branche," in Data Science and Artificial Intelligence, vol. 1942, in: C. Anutariya, M.M. Bonsangue (Eds.), Communications in Computer and Information Science, 1942, Springer Nature Singapore, Singapore, 2023, pp. 157–169, doi:10.1007/978-981-99-7969-1_12.
[3] T. Doan Thi Kieu, T. Ngo Ngoc, K. Kamei, T.T.T. Tran, T.T.N. Nguyen, Applications of bacteriophages in controlling rice bacterial grain rot caused by Burkholderia glumae, CTUJS 13 (3) (2021) 17–22, doi:10.22144/ctu.jen.2021.036.
[4] R. Mizobuchi, S. Fukuoka, C. Tsuiki, S. Tsushima, H. Sato, Evaluation of major rice cultivars for resistance to bacterial seedling rot caused by Burkholderia glumae and identification of Japanese standard cultivars for resistance assessments, Breed. Sci. 70 (2) (2020) 221–230, doi:10.1270/jsbbs.19117.
[5] N. Jungkhun, A.R. Gomes De Farias, J. Watcharachaiyakup, N. Kositcharoenkul, J.H. Ham, S. Patarapuwadol, Phylogenetic characterization and genome sequence analysis of Burkholderia glumae strains isolated in thailand as the causal agent of rice bacterial panicle blight, Pathogens. 11 (6) (Jun. 2022) 676, doi:10.3390/pathogens11060676.
[6] S.A. Mirghasempour, S. Huang, D.J. Studholme, C.L. Brady, A grain rot of rice in iran caused by a xanthomonas strain closely related to X. sacchari, Plant Dis. 104 (6) (2020) 1581–1583, doi:10.1094/PDIS-01-20-0179-SC.
[7] H. Kim, K.S. Do, J.H. Park, W.S. Kang, Y.H. Lee, E.W. Park, Application of numerical weather prediction data to estimate infection risk of bacterial grain rot of rice in Korea, Plant Pathol. J. 36 (1) (2020) 54–66, doi:10.5423/PPJ.OA.11.2019.0281.
[8] Y. Kouzai, C. Akimoto-Tomiyama, A seed-borne bacterium of rice, Pantoea dispersa BB1, protects rice from the seedling rot caused by the bacterial pathogen Burkholderia glumae, Life 12 (6) (2022) 791, doi:10.3390/life12060791.

[9] I. Aflaha, A.J. Chairul, Baharuddin, T. Kuswinanti, Molecular identification of bacteria causing grain rot disease on rice, IOP Conf. Ser.: Earth Environ. Sci. 486 (1) (2020) 012165, doi:10.1088/1755-1315/486/1/012165.
[10] A.A. Darmawan, T. Kuswinanti, A. Asman, Rapid detection of Burkholderia glumae causal agent of grain rot disease in rice seed from Gowa Regency, South Sulawesi using ELISA, IOP Conf. Ser.: Earth Environ. Sci. 807 (2) (2021) 022097, doi:10.1088/1755-1315/807/2/022097.
[11] Y. Wang, H. Wang, Z. Peng, Rice diseases detection and classification using attention based neural network and bayesian optimization, Expert. Syst. Appl. 178 (2021) 114770, doi:10.1016/j.eswa.2021.114770.
[12] J. Chen, D. Zhang, A. Zeb, Y.A. Nanehkaran, Identification of rice plant diseases using lightweight attention networks, Expert. Syst. Appl. 169 (2021) 114514, doi:10.1016/j.eswa.2020.114514.
[13] L.-D. Quach, K.N. Quoc, A.N. Quynh, H.T. Ngoc, Evaluation of the efficiency of the optimization algorithms for transfer learning on the rice leaf disease dataset, IJACSA 13 (10) (2022), doi:10.14569/IJACSA.2022.0131011.
[14] Chiharu Akimoto-Tomiyama, Multiple endogenous seed-born bacteria recovered rice growth disruption caused by Burkholderia glumae, Sci. Rep. 11 (1) (2021) 4177.