Data in Brief 48 (2023) 109108

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

Machine Learning Imagery Dataset for Maize Crop: A Case of Tanzania

Neema Mduma[a,*], Hudson Laizer[b]

[a] The Nelson Mandela African Institution of Science and Technology, Department of Information and Communication Sciences and Engineering, P o Box 447, Tengeru, Arusha - Tanzania
[b] Mbeya University of Science and Technology, Department of Natural Sciences , P o Box 131, Mbeya - Tanzania

ARTICLE INFO

Article history:
Received 5 January 2023
Revised 5 March 2023
Accepted 24 March 2023
Available online 31 March 2023

Keywords:
Maize
Maize Lethal Necrosis
Maize Streak Virus
Leaves
Image

ABSTRACT

Maize is one of the most important staple food and cash crops that are largely produced by majority of smallholder farmers throughout the humid and sub-humid tropic of Africa. Despite its significance in the household food security and income, diseases, especially Maize Lethal Necrosis and Maize Streak, have been significantly affecting production of this crop. This paper offers a dataset of well curated images of maize crop for both healthy and diseased leaves captured using smartphone camera in Tanzania. The dataset is the largest publicly accessible dataset for maize leaves with a total of 18,148 images, which can be used to develop machine learning models for the early detection of diseases affecting maize. Moreover, the dataset can be used to support computer vision applications such as image segmentation, object detection and classification. The goal of generating this dataset is to assist the development of comprehensive tools that will help farmers in the diagnosis of diseases and the enhancement of maize yields thus eradicating the problem of fod security in Tanzania and other parts in Africa.

* Corresponding author's email address and Twitter handle.
E-mail address: neema.mduma@nm-aist.ac.tz (N. Mduma).
Social media: @nakadori (N. Mduma), @hudson_laizer (H. Laizer)

https://doi.org/10.1016/j.dib.2023.109108

Specifications Table

Subject	Applied Machine Learning
Specific subject area	Computer vision techniques for the detection of diseases affecting maize specifically Maize Lethal Necrosis (MLN) and Maize Streak Virus (MSV)
Type of data	Image
How the data were acquired	Data were collected using Samsung Galaxy 01 smartphone cameras with 13-megapixel. Open Data Kit (ODK) application called AdSurv was installed on the smartphones to capture images of maize leaves in the field. Raw data on maize image leaves were classified as either healthy or affected by MLN or MSV. Data collection process involved researchers and farmers while agricultural extension officers and plant pathologists were in charge of the quality check.
Data format	Raw
Description of data collection	Images were collected in the field in the range of six months. The intention was to have the dataset of maize diseases diagnostics with consideration of two identified diseases which are mainly affecting productivity. The names for each disease in the dataset were determined by looking at the caption associated with the maize image sample.
Data source location	• Institution: The Nelson Mandela African Institution of Science and Technology (NM-AIST), Tanzania Agricultural Research Institute (TARI) • City/Town/Region: Arusha • Country: Tanzania
Data accessibility	Repository name: Harvard Dataverse Data identification number: doi:10.7910/DVN/GDON8Q Direct URL to data: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/GDON8Q

Value of the Data

- Maize dataset can be used to train machine learning models for early detection of Maize Lethal Necrosis (MLN) and Maize Streak Virus (MSV) diseases affecting maize productivity.
- Maize image dataset can be used by machine learning researchers to develop technological solutions on addressing problems in the agricultural sector.
- The generated dataset can be used to facilitate various computer vision tasks such as image segmentation, image object detection, and image classification.
- The dataset contains all possible instances and to the best of our knowledge, this is among the biggest publicly accessible dataset on maize in Tanzania.

1. Objective

The aim of generating this dataset is to deliver open, accessible and quality machine learning dataset for crop pests and disease diagnosis based on crop imagery data from Tanzania. The development of beneficial and effective real-world machine learning applications requires localized and well labelled datasets. This imagery dataset provides solution for early disease identification that will help to address the issue of food security in Tanzania. The dataset can be annotated in formats that deliver a wide range of machine learning use cases including classification and object detection tasks. The dataset can also be used by other researchers for modelling the spread of maize diseases which will ultimately help in breeding resistant crop varieties that are necessary for alleviating the problem of food security in Sub-Saharan Africa.

Fig. 1. Image sample of maize leaves (a) healthy (b) MLN (c) MSV.

2. Data Description

This article presents imagery dataset of maize leaves of both healthy and those affected by Maize Lethal Necrosis (MLN) and Maize Streak Virus (MSV) collected from farms in Tanzania. The dataset has a total of 18,148 labeled images with 640 × 480 pixels in jpeg format and a label indicating the name of the image from the image number. In the repository, data were uploaded into 4 separate folders; 1 folder of healthy data submitted in zip format, 1 folder of MLN data submitted in zip format and 2 folders of MSV data submitted in zip format. Also, all folders were named to indicate its corresponding image class. HEALTHY folder contains all images of healthy maize leaves, MLN folder contain images of maize leaves affected by MLN and MSV_1 and MSV_2 folders contain images of maize leaves affected by MSV. Images were separated in different folders to allow easy uploading and downloading of data. Fig. 1 shows samples of maize leaves, both healthy and those affected by MLN and MSV.

3. Experimental Design, Materials and Methods

3.1. Field data collection

The dataset consists of images of maize leaves collected by the Nelson Mandela African Institution of Science and Technology (NM-AIST) and Tanzania Agricultural Research Institute (TARI) located in Arusha, Northern Tanzania. Open Data Kit (ODK) application called AdSurv installed in a Samsung Galaxy 01 with 13 megapixels was used to collect image data. Data collection was done by both researchers and farmers, while data quality check was done by agricultural officers and plant pathologists. Data collection was done in six months, from February to July 2021, and involved farms in Arusha, Kilimanjaro and Manyara regions. The three regions were selected purposely by considering the maize production and diseases prevalence [1].

3.2. Data preprocessing

The collected data from farms were then cleaned, renamed and annotated before uploading in the open access repository. Duplicate images were identified during data cleaning and removed using VisiPics software [2]. However, a paltry remaining duplicates might exist, but the number is small enough to significantly impact training/testing [3]. The dataset might contain distinct images that are not defined to be duplicates but are extremely similar. Table 1 presents number of maize images before and after removing duplicates.

After cleaning the datasets, images were grouped into healthy, MLN, MSV classes and then numbered. LabelMe which is an open source tool was used to annotate preliminary images for object detection (Fig. 2) [4], and the final image annotation was done using web annotation

Table 1
Maize dataset before and after removing duplicates.

Class name	Before removing duplicates	After removing duplicates
Healthy	6531	5118
MLN	6223	3982
MSV	7345	6255

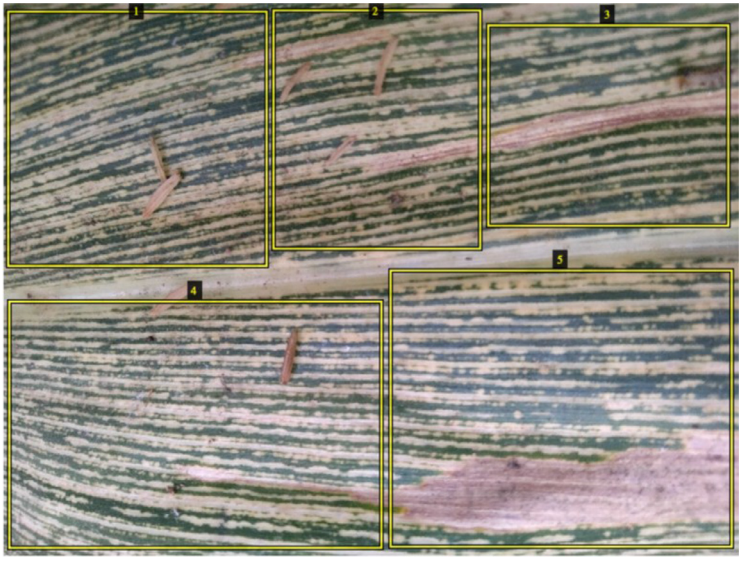

Fig. 2. Sample of the annotated maize leaf.

tool developed by Makerere AI Lab [5]. The final curated images were deposited in the Havard DataVerse open repository [6]. The image quality was of paramount importance for this dataset which would otherwise lead to bias, therefore it was crucial to ensure that the datasets did not contain images captured under extreme low lighting conditions or collected using low quality smartphone cameras.

Ethics Statements

The study does not involve experiments on humans or animals.

Credit Author Statement

Neema Mduma: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization, Supervision, Project administration, Funding acquisition. **Hudson Laizer:** Conceptualization, Methodology, Writing - Review & Editing, Supervision, Project administration.

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.

Data Availability

Maize Dataset Tanzania (Original data) Dataverse.

Acknowledgments

The authors would like to thank the Rockefeller Foundation, Google.org, and Canada's International Development Research Centre for supporting data collection through Lacuna Fund in Agriculture. Also, the authors acknowledge the project partners; Makerere University AI Lab, Namibia University of Science and Technology (NUST) and KaraAgro AI Foundation. This work was also supported by the grant from the Artificial Intelligence for Development in Africa Program, a program funded by Canada's the International Development Research Centre, Ottawa, Canada and the Swedish International Development Cooperation Agency, grant number 109704-001/002. Lastly, we wish to extend our gratitude to Mr. Loyani Loyani, Mr. Mbwana Macheli, Mr. Zablon Msengi, Ms. Alice Karama, and Ms. Irine Msaki for their assistance and technical advice during field data collection and annotation.

References

[1] L. Nyaligwa, H. Shimelis, M. Laing, H. Ghebrehiwot, B. Amelework, Key maize production constraints and farmers' preferred traits in the mid-altitude maize agroecologies of northern Tanzania, South African Journal of Plant and Soil. 34 (2017) 47–53, doi:10.1080/02571862.2016.1151957.
[2] Softonic, Software to detect and remove duplicate pictures. https://visipics.en.softonic.com, 2023 (accessed 10 February 2023).
[3] Q. Chen, J. Zobel, X. Zhang, K. Verspoor, Supervised learning for detection of duplicates in genomic sequence databases, PLoS ONE 11 (2016) 1–15, doi:10.1371/journal.pone.0159644.
[4] Anaconda, Labelme. https://anaconda.org/conda-forge/labelme, 2023 (accessed 10 February 2023).
[5] Makerere AI Lab, Web annotation tool. https://github.com/AI-Lab-Makerere/web-annotation-tool, 2023 (accessed 27 February 2023).
[6] N. Mduma, H. Laizer, L. Loyani, M. Macheli, Z. Msengi, A. Karama, I. Msaki, S. Sanga, Maize dataset Tanzania, Havard Dataverse (2022), doi:10.7910/DVN/GDON8Q.