2024-10

Dataset of Virginia Flue-cured Tobacco Leaf images based on stalk leaf position for classification tasks: A case of Tanzania

Nguleni, Faith

Elsevier

https://doi.org/10.1016/j.dib.2024.110817

Data in Brief 56 (2024) 110817

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

Dataset of Virginia Flue-cured Tobacco Leaf images based on stalk leaf position for classification tasks: A case of Tanzania

Faith Nguleni[a,b,*], Devotha Nyambo[b], Jacob Lisuma[c], Shubi Kaijage[b]

[a] Tanzania Public Service College, P.o Box 329, Tabora, Tanzania
[b] The Nelson Mandela African Institution of Science and Technology, P.o Box 447, Tengeru, Arusha Tanzania
[c] Tobacco Research Institute of Tanzania, P.o Box 431, Tumbi, Tabora, Tanzania

ARTICLE INFO

Article history:
Received 3 July 2024
Revised 22 July 2024
Accepted 5 August 2024
Available online 10 August 2024

Dataset link: Tobacco leaves dataset (Original data)

Keywords:
Grade labels
Plant
Nicotiana tabacum
Factored tobacco

ABSTRACT

Nicotiana tabacum is a kind of plant cultivated for its leaves used for manufacturing medicine and cigarettes. With the common name, the Tobacco plant is grown in many countries including China, Indonesia, Malawi and Tanzania just to mention a few. Literatures suggest a technical gap in the proper identification of grade labels for various parts of the plant. In addition, manual grading has resulted in various gaps and biases. To mitigate this, a data-driven grading solution is necessary. However, relevant datasets to train grade classifiers from various countries become of the essence. This article presents images concentrated on tobacco leaf plant position namely *Leaf position* which normally carries 23 grade labels. Due to high rainfall which swiped away the applied fertilizer on the tobacco plants in the farms, we failed to get images of one grade. Therefore, this research could capture and label 22 grade labels. Images of tobacco leaves based on the tobacco plant position were collected in Tanzania through participatory community research. Canon 5D mark III cameras with 100 mm micro lens were used to take pictures of tobacco leaves based on the tobacco plant position. Domain experts were used for image labelling and cleaning according to tobacco grade labels identified in Tanzania. The dataset carries 49,779 images, which can be used to develop machine learning models for tobacco leaf grade label identification. The

* Corresponding author.
E-mail address: ngulenif@nm-aist.ac.tz (F. Nguleni).

https://doi.org/10.1016/j.dib.2024.110817

collected dataset can be used to train models and enhance the performance of pre-trained models in any country of interest.

Specifications Table

Subject	Applied Machine Learning
Specific subject area	Computer vision techniques for classifying the grade labels of tobacco leaves based on stalk leaf position
Type of data	Image
Data collection	Data were collected using two Canon EOS 5D Mark III cameras with 100 mm micro lens. Capture one software was installed in two computers which was used to capture tobacco leaves pictures. This software was connected between the cameras and the computers using firewire cables allowing the images to directly display on the laptop computers. The tobacco leaf position was the region of interest as are the higher leaves position of tobacco with stronger flavour and good grade labels as the leaves gets more time to mature and develop their flavour [1]. Researchers, farmers and tobacco leaves experts participated in the data collection.
Data format	Raw
Description of data collection	Tobacco leaves images were collected in the field for two months from February 2024 to April 2024. Tobacco leaves images were collected in agricultural season 2023/2024 for two months from February 2024 to April 2024 as it was the only time to collect data because tobacco plant is grown just in one season in the year in Tanzania and as crops are to be sold and exported to other countries [2]. Tobacco leaf position namely *LEAF POSITION* was taken into consideration when compiling the dataset. By examining the caption for tobacco leaves image sample, the names of each grade label within leaf position in the dataset were identified.
Data source location	Tanzania Tobacco Board (TTB), Tobacco Research Institute of Tanzania (TORITA) City/Town/Region: Tabora Country: Tanzania
Data accessibility	Repository name: Harvard Dataverse Direct URL to data: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/TTPLFT

1. Value of the Data

- To develop automated tobacco leaves grading systems that use Machine Learning models for improved accuracy and efficiency during grading processes in the tobacco marketplaces.
- Can be used by developers of state-of-the-art artificial intelligence solutions in learning institutions.
- Our dataset comprises 22 grade labels which is considerably more than most of existing tobacco leaves datasets. For example, the study [3] had a dataset with 9 tobacco leaf grade labels, also the dataset used [4] had 6 tobacco leaf grade labels.
- Most tobacco leaf datasets combined all the leaves from different tobacco plant positions, but our dataset separated the leaves based on stalk leaf position as it is very crucial when developing grade labelling systems.
- To the best of our knowledge, this is the biggest and only tobacco leaves dataset in Tanzania which has every potential case.

2. Background

The main purpose of generating a Virginia flue-cured tobacco leaf dataset is to help researchers easily access the tobacco leaf dataset that will help in research activities, especially in Africa, as the dataset will be accessible in open-source dataset repositories. In the context of Tanzania, there is no publicly available dataset of Virginia flue-cured tobacco stalk leaves positions to the best of our knowledge. Our dataset consists of 49,779 images while most existing tobacco leaves datasets consist of fewer images example the study [5] used a dataset with 21,113 images, the study [3] collected a dataset with 11,849 images and the study [4] used a dataset with 22,322 images of which all the authors did not consider the stalk leaves positions.

The dataset will be used by researchers in developing machine learning and artificial intelligence solutions. These solutions will address the agricultural challenges and assist in crop management activities in Tanzania and other areas in Africa. The samples produced from the Virginia flue-cured tobacco leaves dataset can be used to provide different kinds of machine learning use cases, such as classification and object detection. Also, this dataset can be used by different researchers in simulating and analysing the grading process of tobacco leaves; hence, it will help to create consistent models aimed at economic development and agricultural improvement in Sub-Saharan Africa.

3. Data Description

The presented dataset is for tobacco leaves based on tobacco leaf position from tobacco farmers in Tanzania. A total of 49,779 images with the grade label indicating the names of each grade label with their specific total number of images are presented in Table 2. The dataset has 22 folders, and each folder is named after the grade label name.

Tobacco has similar fundamental principles that govern the grading of tobacco leaves in all nations that produce tobacco. However, because each nation has its own distinct set of regulations and evaluation procedures, there is a wide range of grades given to tobacco leaves [4]. The labels given to various grades of tobacco leaves in different countries that produce tobacco products vary. This suggests that different geographical areas use different terms to define the quality levels of tobacco leaves.

A good example is based on stalk leaf position; in China, the upper tobacco leaf is labelled with the letter **B** for example, **B2F** defines an upper part of tobacco leaf that has a quality of 2 and is orange in colour [4]**,** in Indonesia is labelled by three letters named **TNG** [6] while in Tanzania the upper tobacco leaf is labelled by letter **L** for example **L2O** defines an upper part of tobacco leaf that has a fine quality of 2 and is orange colour [2].

The grade label images in our dataset follow the Tanzanian standards of grade labelling, whereby the name starts with a letter that indicates the leaf position of the leaf on the plant of the tobacco, followed by a number that indicates the quality of the tobacco leaf and lastly followed by a letter that indicates the colour of the tobacco leaf. Additionally, only three colours are considered which include Orange (O), Lemon (L) and Mahogany (R) [2]. For other countries to use this dataset, naming conversion can be performed to match their grade naming systems.

Tobacco leaf grade labels can have the letter F after the basic grade labels, which is used only when the tobacco has a special factor such as maturity spots or fully matured or fully orange coloured. Fig. 1 shows a sample of tobacco leaves grade labels (L1L, L5R and L1OF) with the following meaning: L1L (plant position = Leaf (L), Choice quality = 1, Colour = Lemon(L)), L5R (plant position = Leaf (L), Low quality = 5, Colour = Mahogany (R)) and L1OF (plant position = Leaf (L), Choice quality = 1, Colour = Orange (O), F = Special Factored tobacco). Table 1 describes the Tanzanian grade labels naming structure.

One grade label named L1R was not captured during this round of data collection. The plan is to collect it in the next round of data collection. Although the grade label was not collected now, it does not affect the quality of other grades that have been collected.

Fig. 1. Image sample (a) L1L (b) L5R (c) L1OF.

Table 1
Tanzanian grade labels naming structure.

Grade label name example: L		1		O		F	
Section 1 (Plant position)		Section 2 (Leave quality)		Section 3 (Colour)		Section 4 (Special factor)	
Symbol	Meaning	Symbol	Meaning	Symbol	Meaning	Symbol	Meaning
C	Cutters position	1	Choice quality	L	Lemon	F	Special factors (Maturity spots OR full maturity OR full orange colour)
L	Leaf position	2	Fine quality	O	Orange		
M	Thin Leaf position	3	Good quality	R	Mahogany		
X	Lugs and Primings position	4	Fair quality				
		5	Low quality				

4. Experimental Design, Materials and Methods

4.1. Field data collection

Tobacco leaf images based on leaf position was collected in collaborative research by the Nelson Mandela African Institute of Science and Technology (NM-AIST) located in Arusha, Tobacco Research Institute of Tanzania (TORITA) located in Tabora Tanzania and Tanzania Tobacco Board (TTB) in Tanzania. Canon 5D Mark III cameras were used, and for them to communicate with a computer, capture one software was installed in all computers that were used to capture the tobacco leaves images. In the process of capturing tobacco leaves images, a flat table was set with a camera on top at a 90-degree angle. The camera was set at approximately, 2m above the table. The tables were covered with a sack to maintain a consistent background colour. As an attempt to control lighting conditions, white tents were used as light diffusers. This set was used to collect all the images in all the data collection centres regardless of the regions. This was done to maintain the consistency and accuracy of the dataset.

Farmers were prepared three months before the visit for data collection so that they can prepare the tobacco leaves in various grades. The data collection was done by researchers, farm-

Table 2
Tobacco grade labels images before and after deleting images that were captured more than once and those that were not clear.

Grade labels (class name)	Before deleting images that were captured more than once and those that were not clear	After deleting images that were captured more than once and those that were not clear
L1L	4638	3570
L10	3746	1413
L10F	3674	2021
L2L	6783	4914
L20	2989	2707
L20F	4893	3421
L2R	178	32
L3L	9848	7093
L3O	7937	5566
L3OF	3748	3615
L3R	580	262
L4L	2890	1594
L40	4892	3009
L4R	789	348
L5L	2693	1761
L50	3682	1890
L5R	2832	1341
LG	1738	865
LK	1987	605
LLV	4012	2106
LND	478	187
LOV	2564	1459

ers and domain experts from Tanzania Tobacco Board (TTB). In Tanzania, there are government regions and tobacco regions [2]. The tobacco leaf images were collected for two months from February 2024 to April 2024 where four different tobacco regions were involved and within each tobacco region, five farms were involved in image collections. The tobacco regions involved in this study were Tabora municipal, Uyui, Urambo and Kaliua. Reasons for choosing the named regions was due to the high tobacco production compared to other tobacco regions in Tanzania [7]. In each region and from every farmer, all grade labels were able to be collected in different quantities. Due to this, the geographical diversity was minimal and thus was not considered to have a substantial impact.

4.2. Data pre-processing

The collected tobacco leaf images were grouped from different tobacco regions and the images were renamed as (img-x), where x is a sequential number from 1 to n, whereby n is the number of images per grade label (class). Cleaning of the images was another task whereby any image that was not clear and captured more than once was deleted from the image list [8]. Table 2 presents the number of grade labels of tobacco leaf images based on a leaf position before and after deleting images that were captured more than once and those that were not clear.

Limitations

The presented dataset was supposed to carry 23 grade labels but due to high rainfall during the tobacco agricultural season 2023/2024 which carried away the fertilizer on the tobacco plants in farms, we managed to collect 22 grade labels and one grade label named L1R did not get any image during data collection.

Ethics Statement

Human experiments were not involved in this study.

CRediT Author Statement

Faith Nguleni: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Project administration, Writing – review & editing, Visualization, Funding acquisition; **Jacob Lisuma**: Supervision, Project administration; **Devotha Nyambo**: Conceptualization, Writing – review & editing, Supervision; **Shubi Kaijage**: Writing – review & editing, Supervision.

Data Availability

Tobacco leaves dataset (Original data) (Dataverse).

Acknowledgments

Special thanks go to Tobacco Research Institute of Tanzania (TORITA) for supporting data collection through funding transport fuel during data collection activities. The authors also acknowledge Tanzania Tobacco Board (TTB) for allowing tobacco leaves experts to participate and assist in identifying the required tobacco leaves during the data collection process. Lastly, we wish to extend our gratitude to Mr. Edwin Kambo, Mr. Elimboto Muna, Mr. Charles Mezza, Mr. Albert Chale, Mr. Josephat Tesha, Mr. Deusdedis Mabula and Mr. Innocent Ilomo, for their assistance and technical advice during data collection.

Declaration of Competing Interest

In this reported work, there are no known competing financial interests.

References

[1] L. Zhang, et al., Metabolic profiling of tobacco leaves at different growth stages or different stalk positions by gas chromatography–mass spectrometry, Ind. Crops Prod. 116 (February) (2018) 46–55, doi:10.1016/j.indcrop.2018.02.041.
[2] AnonTTB, Tanzania Tobacco Board, 2023 https://www.tobaccoboard.go.tz/tobacco-growing-areas (Accessed 6 August 2023).
[3] M. Xu, J. Gao, Z. Zhang, X. Guo, Multi-channel and multi-scale separable dilated convolutional neural network with attention mechanism for flue-cured tobacco classification, Neural Comput. Appl. 35 (21) (2023) 15511–15529, doi:10.1007/s00521-023-08544-7.
[4] M. Lu, et al., Intelligent grading of tobacco leaves using an improved bilinear convolutional neural network, IEEE Access 11 (May) (2023) 68153–68170, doi:10.1109/ACCESS.2023.3292340.
[5] X. Xin, H. Gong, R. Hu, X. Ding, S. Pang, Y. Che, Intelligent large-scale flue-cured tobacco grading based on deep densely convolutional network, Sci. Rep. 13 (1) (2023) 11119, doi:10.1038/s41598-023-38334-z.
[6] A. Harjoko, A. Prahara, T.W. Supardi, I. Candradewi, R. Pulungan, S. Hartati, Image processing approach for grading tobacco leaf based on color and quality, Int. J. Smart Sens. Intell. Syst. 12 (1) (2019) 1–10, doi:10.21307/ijssis-2019-010.
[7] K. Domitius, Impact of Tobacco Curing on the Environment Andsocio-Economic Aspects in the Urambo District, 2021 [Online]. Available: https://dspace.nm-aist.ac.tz/handle/20.500.12479/1316.
[8] N. Mduma, J. Leo, Dataset of banana leaves and stem images for object detection, classification and segmentation: a case of Tanzania, Data Br. 49 (2023) 109322, doi:10.1016/j.dib.2023.109322.