Non-destructive insect infestation detection of pear fruits by X-ray imaging and deep learning.

A Data Management Plan created using DMPonline.be

Creator: Jiaqi He

Affiliation: KU Leuven (KUL)

Funder: Fonds voor Wetenschappelijk Onderzoek - Research Foundation Flanders (FWO)

Template: FWO DMP (Flemish Standard DMP)

Grant number / URL: 1SH8Q24N

ID: 206751

Start date: 01-11-2023

End date: 01-11-2027

Project abstract:

Pear fruit is the most economically valuable fruit for Belgium. Unfortunately, pest infestation especially by the codling moth (Cydia pomonella) regularly occurs during fruit growth. The insect feeding often occurs within the fruit without showing obvious external symptoms which fail to be detected by current commercial quality grading systems based on external properties of fruits. To deal with this, pears are sampled and dissected for visually inspection. If several pests are found, the whole batch will be discarded despite of containing a significant amount of healthy fruit resulting in food waste. Moreover, not all fruit is inspected, thus infested fruit with pathogens can reach consumers triggering food safety issues. Additionally, potential exporting markets have stringent quarantine measures for invasive organisms. If one pest is spotted, the whole consignment would be rejected leading to huge economic loss.

Therefore, the aim of this PhD project is to develop a non-destructive insect infestation detection system to identify pears with various damages caused by all stages of codling moth by using X-ray imaging and deep learning. Pears infested with different stages of larvae will be prepared and characterized in 3D, different data augmentation strategies are used to provide sufficient data for developing deep learning models. The method will be tested on the real data collected from growers and will be explored for pest infestation of other fruits and vegetables.

Last modified: 24-04-2024

Non-destructive insect infestation detection of pear fruits by X-ray imaging and deep learning.
FWO DMP (Flemish Standard DMP)

1. Research Data Summary

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.

				Only for digital data	Only for digital data	Only for digital data	Only for physical data
Dataset Name	Description	New or reused	Digital or Physical	Digital Data Type	Digital Data format	Digital data volume (MB/GB/TB)	Physical volume
		Please choose from the following options: • Generate new data • Reuse existing data	*Please choose from the following options:* • Digital • Physical	*Please choose from the following options:* • Observational • Experimental • Compiled/aggregated data • Simulation data • Software • Other • NA	*Please choose from the following options:* • .por, .xml, .tab, .csv,.pdf, .txt, .rtf, .dwg, .gml, … • NA	*Please choose from the following options:* • <100MB • <1GB • <100GB • <1TB • <5TB • <10TB • <50TB • >50TB • NA	
Samples	pear fruit infested with and without codling moth	Generate new data	Physical	Experimental	NA	NA	1000-1500 fruit
X-ray medical CT scans	pear fruit infested with and without codling moth	Generate new data	Digital	Experimental	.dcm	2.5GB/30 fruit <100GB	
X-ray micro CT scan	pear fruit infested with and without codling moth	Generate new data	Digital	Experimental	.tif	22GB/fruit <5TB	
X-ray radiography	pear fruit infested with and without codling moth	Generate new data	Digital	Experimental	.tif	50MB/fruit <100GB	
CT image processing			Digital	Software (Avizo, Python, Matlab, ImageJ)	.tif, .am, .m, .png, .csv	<100GB	
Scripts and models			Digital	Software (Python, Matlab)	.py, ipynb,.m, .pt	<500MB	
Photos	images of cut-open fruit	Generate new data	Digital	Experimental	.jpg	100KB/fruit <1GB	

If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:

No

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? Describe these issues in the comment section. Please refer to specific datasets or data types when appropriate.

- No

Will you process personal data? If so, briefly describe the kind of personal data you will use in the comment section. Please refer to specific datasets or data types when appropriate.

- No

Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)? If so, please comment per dataset or data type where appropriate.

- No

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material/Data transfer agreements/ research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.

- No

Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.

- No

2. Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future (e.g., in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, Codebook.tsv etc. where this information is recorded).

1. X-ray (micro-)CT images: automatically generated .txt files containing the scanning and reconstruction settings will be included to the folders containing the projection and reconstructed (micro-)CT images, respectively. Additionally, a template based Word document will be added for each scan, containing information on the scanned fruit material, such as the fruit moth stages, procedure of infestation for fruit.
2. Medical CT data: automatically generated .txt files containing the scanning and reconstruction settings will be included to the folders containing the reconstructed CT images. Additionally, a template based Word document will be added for each scan, containing information on the scanned fruit material, such as the fruit moth stages, procedure of infestation for fruit.
3. Radiographies: the radiographies together with the flatfield and darkfield image will be collected in folders. Additionally, a template based Word document will be added for each scan, containing information on the scanned fruit material, such as the fruit moth stages, procedure of infestation for fruit.
4. Models: all the scripts for a neural network will be managed by version control in Gitlab. Also, the trained models will be available there. Additionally, a ReadMe file will be added discussing the application of the model and the choice for certain model parameters.

Will a metadata standard be used to make it easier to find and reuse the data? If so, please specify (where appropriate per dataset or data type) which metadata standard will be used. If not, please specify (where appropriate per dataset or data type) which metadata will be created to make the data easier to find and reuse.

- No

No metadata standard is available for the used and generated data. The main effort will be to align metadata and documentation with group common practices through a metadata portal RDIS designed by MeBioS group.
I will use guidelines for documentation of used programming languages (python, Matlab) with provided metadata infrastructure on Gitlab for version control and documentation. ReadMe files and documentation will be provided as explained in the previous entry.

3. Data storage & back-up during the research project

Where will the data be stored?

The data will be stored on a network drive of KU Leuven. The folder is open for the group members and is secured and backed-up by the ICTS service of KU Leuven. The model's code will be available at the university GitLab with restricted access to the research group. Published data will be available according to the publisher standards. All data are locally saved on one of the hard disks on the lab's personal desktop computer.

How will the data be backed up?

The general network drive is maintained by the ICTS service of KU Leuven with automatic daily back-up and mirror procedures. Automatic backup of the drive is further secured on university One Drive cloud with regular automatic backup as per the group guidelines. The backup of code is secured by Gitlab versioning with every commit.

Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely.
If no or insufficient storage or backup capacities are available, then explain how this will be taken care of.

- Yes

Yes, KU Leuven provides sufficient storage and back-up capacity during and after the project. A dedicated storage share will be made for the project on which the collaborators will work jointly and store data files. For the expected large volume of 3D image data, we will rent separate large volume storage of the ICTS service.

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

The network drive of the shared folder is secured by the ICTS service of KU Leuven with a mirror copy. Only specific lab members will have access to the shared folder. Unauthorized persons do not have access to this system. GitLab repository can be managed only by its author and responsible ICTS personnel and is secured by restricted access only to the lab members.

What are the expected costs for data storage and backup during the research project? How will these costs be covered?

Type 1 server back-end storage with mirror backup for the project share folder will cost 519 Euro per Tb per year. The large volume storage will cost 156.60 Euro per Tb per year. Costs will be covered by the project consumables budget.

4. Data preservation after the end of the research project

Which data will be retained for at least five years (or longer, in agreement with other retention policies that are applicable) after the end of the

project? In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).

All data obtained during this FWO project will be retained for the expected 5 year period. Even after this period, the data will remain available for lab members of the MeBioS group.

Where will these data be archived (stored and curated for the long-term)?

Essentail data will be stored on the university's central server large volume storage. Essential model files will remain in the Gitlab.

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?

Cost of the large volume storage will be € 156,60 per TB and year. We anticipate that we will need 5 TB for 5 years to keep the essential data available. This will amount to 3915 euro and will be covered by the general budget and successive projects of the participating groups.

5. Data sharing and reuse

Will the data (or part of the data) be made available for reuse after/during the project? In the comment section please explain per dataset or data type which data will be made available.

- Yes, in a restricted access repository (after approval, institutional access only, …)

All data will be published and made available after the end of the project. Data with valuable IP will be protected prior to publication. The MeBioS group is implementing a web-based platform for sharing of CT data which can be used to share the 3D image data.

If access is restricted, please specify who will be able to access the data and under what conditions.

All data will be available without restrictions to all the lab members. Published data will be
available for everybody with access to the publication as per publisher's rules. Metadata of all data will be available on the MeBioS metadata portal RDIS with a possibility to request the complete dataset for research purposes.

Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)? Please explain in the comment section per dataset or data type where appropriate.

- No

Where will the data be made available? If already known, please provide a repository per dataset or data type.

- In a restricted access repository
- Upon request by mail
- Other (specify):

The data will be stored and be available for lab members using a shared network drive and large volume storage provided by the university. Metadata of all data will be available on the MeBioS metadata portal with a possibility to request the complete dataset for research purposes. Model files will be available with restricted access to the lab members on Gitlab.

When will the data be made available?

Upon publication of the research results

Which data usage licenses are you going to provide? If none, please explain why.

For the data, Creative Commons Attribution-NonCommercial-ShareAlike (CC-BY-NC-SA) will be provided.
For the code, Common Development and Distribution License (CDDL-1.0) will be provided.

Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, you have the option to provide it in the comment section.

- No

What are the expected costs for data sharing? How will these costs be covered?

Expected data sharing costs are minimal and covered by university services.

6. Responsibilities

Who will manage data documentation and metadata during the research project?

Jiaqi He - FWO fellow

Who will manage data storage and backup during the research project?

The FWO fellow (Jiaqi He) will be responsible to store the data on the appropriate accommodation provided by KU Leuven. The ICTS service of KU Leuven is responsible for the back-up of the network drives at KU Leuven. The folders will be managed by the supervisors.

Who will manage data preservation and sharing?

Promotors and ICTS.

Who will update and implement this DMP?

Jiaqi He -FWO fellow