Plan Overview

A Data Management Plan created using DMPonline.be

Title: Efficient AI based internal quality inspection of horticultural products using spectral X-ray imaging

Creator: Hugo Li

Affiliation: KU Leuven (KUL)

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

Template: FWO DMP (Flemish Standard DMP)

Project abstract:

Horticultural products are susceptible to significant food loss due to quality degradation during the postharvest process. While external defects can be readily identified using visual inspection and automated sorting lines, detecting internal disorders necessitates advanced imaging techniques for non-destructive quality assessment. Recent research has demonstrated the potential of artificial intelligence (AI) in detecting internal quality issues in horticultural products through X-ray imaging. However, this approach encounters several challenges including inadequacy of annotated data, limited algorithm generality, unsatisfactory scan quality, and slow image processing speeds. To tackle these obstacles, this PhD project proposes leveraging efficient semi-supervised and unsupervised learning frameworks as well as exploring novel spectral X-ray imaging to fully harness AI capabilities in automatically detecting internal quality issues in fruit and vegetables. Moreover, the project seeks to establish a comprehensive and automated quality control process for the food industry, thereby mitigating food loss and ensuring product quality across the supply chain.

ID: 211169

Start date: 01-11-2024

End date: 01-11-2028

Last modified: 27-01-2025

Efficient AI based internal quality inspection of horticultural products using spectral X-ray imaging 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	Apples and pears at harvest, after 3 months of storage and after 8 months of storage	Generate new data	Physical	Experimental data			~2000 samples
Pome fruit CT	Kanzi apple & Conference pear spectral CT scanned at harvest, after 3-month storage, and after 8-month storage	Generate new data	Digital	Experimental data	.IMA .dcm .tiff .h5	<10TB	
Pome fruit simulated radiograph	Simulated 2D spectral radiographs of pome fruit CT using Astra toolbox	Generate new data	Digital	Simulation data	.tiff	<100GB	

Pome fruit real radiograph	Real 2D radiographs of pome fruit scanned by line scanner	Generate new data	Digital	Experimental data	.tiff	<100GB	
Braeburn storage CT	CT scans of the 2021-2023 Braeburn storage experiment	Reuse existing data	Digital	Experimental data	.IMA	<1TB	
Deep learning models	Deep learning model with trained weight	Generate new data	Digital	Software	.pth	<100GB	

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:

Braeburn storage CT: stored on Large Volume Storage Collection (L drive) of KU Leuven, managed by Postharvest group, Department of Biosystems, KU Leuven

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).

CT data:

- **Metadata** about scan settings (e.g., KVP, reconstruction parameters, and pixel size) is automatically generated by the scanner and stored in each DICOM file. A .txt copy of the DICOM attributes will be generated to enhance accessibility.
- **MATLAB and Python scripts** for preprocessing spectral CT data will be managed by the version control system in GitLab, complete with detailed comments and a well-structured README.md file. The README will outline the preprocessing procedure and provide instructions for use.

Radiograph data:

- **Metadata** about scan settings will be automatically generated as a .txt file for real 2D radiographs of pome fruit scanned by line scanner.
- **Scripts** for generating simulated radiographs will be managed by the version control system in GitLab, complete with detailed comments and a well-structured README.md file. The README will outline the simulation procedure, explain the simulation parameters, and provide instructions for use.

Models:

- **Trained models and scripts for model training and testing** will be managed by the version control system in GitLab, complete with detailed comments and a well-structured README.md file. The README will outline the training and testing procedure, hyperparameters, and instructions for use.

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

For radiograph and CT data, structured metadata elements are automatically generated by scanners, ensuring a consistent metadata standard. Additionally, a structured metadata schema will be created using the KU Leuven MANGO platform to specify sample information (e.g., fruit batch, season).
For Python and MATLAB scripts, no strict standards currently exist for writing README.md files. Therefore, user-friendly documentation will be created to ensure ease of understanding.

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

Where will the data be stored?

Scripts and models will be managed using GitLab for version control and stored locally in the project owner's OneDrive folder.
Active research data will be stored on the network drive for quick access, while **archived data** will be compressed and stored on the MANGO platform. Both storage solutions are provided and managed by KU Leuven's ICTS.
Published data will be made available following the publisher's standards.

How will the data be backed up?

Backups of the **model and script** are conducted with every commit and push to the GitLab version control system. Additionally, daily backups from OneDrive to the group's network drive are automatically performed, following the research group's guidelines and utilizing SyncBackFree software.
Backups of **active and archived data** on the network drive are managed by KU Leuven's ICTS and are performed automatically multiple times per day.

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

Although spectral CT data volumes are significantly larger than those of conventional CT data, the research group ensures sufficient storage and backup capacity both during and after the project by leveraging multiple storage solutions (OneDrive, network drive, and MANGO) provided by KU Leuven's ICTS. All solutions are scalable, allowing capacity to be expanded as needed.

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

The network drive and MANGO platform are secured by KU Leuven's ICTS service. Access to the owner's folders is restricted to designated lab members and external partners. Unauthorized individuals are not granted access to these systems. GitLab repositories are managed solely by the author, with access restricted to lab members for added security.

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 euros per TB per year. The MANGO data platform for archiving data will cost 35 euros 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 the data obtained during this FWO project will be retained for 5 years after the end of the project. 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)?

During the project, inactive data will be stored on the MANGO platform provided by KU Leuven's ICTS. At the conclusion of the project, data will be transferred and archived in the upcoming COLD STORAGE platform, specifically designed for data archiving and managed by ICTS of KU Leuven. Scripts and trained models will be kept in GitLab.

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

The cost of the MANGO platform will be 35 euros per TB and year. It is anticipated that 10TB for 5 years will be needed for data preservation. This will amount to 1750 euros 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 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 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 the publisher's rules. Metadata of all data will be available on the MeBioS metadata portal with the 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 available for lab members using a shared network drive and data management platform provided by the university.
Metadata of all data will be available on the MeBioS metadata portal with the possibility to request the complete dataset for research purposes.
Neural network 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?

Junyan Li - FWO fellow

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

The FWO fellow (Junyan Li) 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?

Junyan Li - FWO fellow