Proximal vegetable diseases datasets

1 Description

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset was created for the task of exploring the feasibility and potential of deep learning neural network architectures such as graph neural networks and to compare their performances to conventional neural networks, in an optic of biovigilance.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

This dataset is an aggregate of two distinct data sources. The first half is composed of plant and fungal observations of carrot, onion and lettuce from distinct fields in 2021. These observations of plant characteristics and fungal symptoms were made by Dr. Carisse and her team on behalf of Agrigulcure and Agri-Food Canada (AAFC). The second half of the dataset is composed of weather-related data obtained through field weather stations collected by Dr. Carisse's team and publicly accessible weather stations from Environment and Climate Change Canada. The aggregation was made by Dr. Lord's team at AAFC.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

This dataset creation was funded through AAFC project J-002366.

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

There are three types of instances in the dataset. The first type is a tabular representation of individual plant characteristics and observations of fungal infection symptoms on their leaves.

This data is associated to a plant ID and a farm ID and a specific day of the year (DOY), as to represent the fluctuations in the various attributes of each plant across spatial and temporal dimensions.

The second data type is also tabular and represents daily weather-related attributes (raw and processed) for the entire plant growth season and across all crop fields observed in the first data type presented above. This data is also farm ID - dependent and has a spatial and temporal component.

The third data type is a list of relative distances (in meters) between each plants pair (per their respective plant ID) in each field sampled. This allows for the tracking of spatial proximity between plants.

The dataset contains data for three distinct crops (carrot, lettuce and onion) collected during the 2021 summer season.

How many instances are there in total (of each type, if appropriate)?

For the onion data, a total of 922 plant observations (rows) are present, as well as 98 days of weather data across 5 fields (with fluctuating dates between them). 3000 distance pairs have been recorded.

For the lettuce data, a total of 588 plant observations (rows) are present, as well as 98 days of weather data across 4 fields (with fluctuating dates between them). 2400 distance pairs have been recorded.

For the carrot data, a total of 519 plant observations (rows) are present, as well as 99 days of weather data across fields (with fluctuating dates between them). 1800 distance pairs have been recorded.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

This dataset is the complete dataset. The dataset raw data has been manually curated to avoid inconsistencies and statistical analyses have been performed ad hoc to insure its reliability.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

The biological data (plants characteristics and fungal infection symptoms) contains features for each plant at a given day, such as the number of healthy and unhealthy leaves and the growth stage.

The base weather data consists of 8 distinct attributes :

- Temperature (Celsius)
- Dewpoint (Celsius)
- Relative humidity (%)
- Solar radiation (W/m$\hat{2}$)
- Rain (millimeters)
- Wind speed (m/s)
- Gustspeed (m/s)
- Wind direction (degree)

Additional daily weather attributes have been computed and are presented in depth in the "Additional Data" section at the end of the document. Growing degree days (GDDs), used to estimate the growth and development of plants during the growing season, have also been computed.

In total, 129 daily weather attributes are contained within the dataset (see table in section 2 - Additional data).

The distance pairs between plants have been processed using the original GPS coordinates. Each line in this text file consist of 4 distinct comma-separated values:

1. Farm ID
2. Plant ID (origin*)
3. Plant ID (destination*)
4. Distance (meters) between the two plant ID

   *Note that no direction is implied by the numbering of the plant ID nor with the origin or destination identification.

Is there a label or target associated with each instance? If so, please provide a description.

Each crop has three distinct labels, each representing an incidence score for a specific fungal species. The identification was based on visual observation and, if needed, by further analytical lab analysis.

The onion dataset has labels for the following fungal incidences:

1. **Botrytis squamosa**. This label represent an incidence score between 0 and 4, where 0 = No observed spots, 1 = 1 to 5 spots, 2 = 6 to 10 spots, 3 = 11 to 20 spots, 4 = 21 spots and more.
2. **Peronispora destructor**. This label represent an incidence score between 0 and 4, where 0 = No observed spots, 1 = 1 to 2 spots, 2 = 3 to 5 spots, 3 = 6 to 9 spots, 4 = 10 spots and more.
3. **Stemphylium vesicarium**. This label represent an incidence score between 0 and 4, where 0 = No observed spots, 1 = 1 to 2 spots, 2 = 3 to 5 spots, 3 = 6 to 9 spots, 4 = 10 spots and more.

The lettuce dataset has labels for the following fungal incidences:

1. **Sclerotinia sclerotiorum**. This label represent a binary incidence score where 0 = absent and 1 = present.
2. **Botrytis cinerea**. This label represent a binary incidence score where 0 = absent and 1 = present.
3. **Bremia lactucae**. This label represent an incidence score between 0 and 4, where 0 = No spots, 1 = 1-10%, 2 = 11-35%, 3 = 26-50%, 4 = 50% and more.

   *Note that for each of those species, the incidence score correspond to observed presence of the corresponding disease symptoms. Assignation was done .

The carrot dataset has labels for the following fungal incidences:

1. **Cercospora carotae**. This labels represent an incidence score between 0 and 4, where 0 = No spots, 1= 1-5%, 2=15%, 3=25%, 4=50% and more.

2. **Sclerotinia sclerotiorum**. This label represent a binary incidence score where 0 = absent and 1 = present.

3. **Alternaria dauci**. This label represent a binary incidence score where 0 = absent and 1 = present.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

All instances are complete due to the preprocessing done.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

All three datatypes are associated through their shared Farm ID, Plant ID and sample date attributes.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

Data splits will depend heavily on the objective of the user and the deep learning architecture employed. There is a case to be made for splitting the data based on Farm ID, as to keep certain farms entirely "unseen" by the model until testing. This can be arduous however, depending on the training method because of the limited number of farms available for each crop.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

No error or source of noise is known at the present.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

All identifiable information within the dataset has been removed.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

Does the dataset relate to people? If not, you may skip the remaining

questions in this section.
The dataset is not related to people.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
N/A.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.
No.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.
N/A

Any other comments?

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
The biological data related to plant health was obtained through direct visual inspection in the field by qualified staff. The weather and distance data originate directly from weather stations and GPS data.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?
The biological data was obtained through direct visual inspection on the field by qualified staff. An additional curation was done manually during pre-processing. Weather and distance data were also manually curated during pre-processing.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?
The strategy employed has been to only keep data related to farms that have been observed throughout the growth season (no premature or unexpected stoppage). This has been done to ensure the temporal structure of the dataset.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?
This data was collected by the Government of Canada.

Over what timeframe was the data collected? Does this timeframe

match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The data was collected between June and September 2021. The data was curated between 2022-2023 for the creation of the final dataset.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

No ethical review has been conducted.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

N/A

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

N/A

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

N/A

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Any other comments?

N/A

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

For the biological data (plant health and fungal infection symptoms), the health of individual leaves has been combined to represent the health of the plant. The fungal infection score of the leaf with the most symptoms has been assigned to the entire plant.

The weather data collected were used to determine the dew point, leaf wetness (Gleason model) and growing degree-days at each site for their respective crops. The hourly data was then aggregated into daily weather data and a rolling average was computed with window sizes of 3, 6 and 14 days . Those transformations are based on state-of-the-art research used for fungal growth prediction in the field.

No transformation has been done to the plant distance data.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

The raw data exists within AAFC in Dr. Carisse's and Dr. Lord's research groups.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

The data has been processed using Python. The scripts are accessible from Dr. lord's research group.

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

The dataset is currently being used in an ongoing project with the goal of exploring the feasibility and potential of deep learning neural network architectures such as graph neural networks and to compare their performances to conventional neural networks, in an optic of biovigilance.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

No, the paper regarding this project is yet to be published.

What (other) tasks could the dataset be used for?

This dataset could be used in several tasks related to biovigilance or predictive models using neural networks or other approaches. It is also very well suited to research possibilities requiring a temporal component, such as long-short term memory (LSTM) models.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

No this dataset does not have potential for harm.

Are there tasks for which the dataset should not be used? If so, please provide a description.

No.

Any other comments?

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub) Does the dataset have a digital object identifier (DOI)?

Due to its small file size (less than 2 MB), the dataset will be distributes to third parties directly through email.

When will the dataset be distributed?

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

Any other comments?

Maintenance

Who will be supporting/hosting/maintaining the dataset?

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Is there an erratum? If so, please provide a link or other access point.

No.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

There are no plans for updates at this time.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

N/A

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

There are no plan to update the dataset and as such, not obsolescence risk.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

No there is not. This is a small niche dataset that does not need contributing to.

Any other comments?

2 Additional data

Table 1: Graphical representation of the weather attributes generated.

Weather predictors		**Temporality**	**Measure**	**Unit**
Temperature		Daily, day and night	average, minimum, maximum	°C
	Temperature of the last 6 days	Daily		
	Hours of temperatures <5°C	Daily, day, night and 6 days	Sum	Hours
	Hours of temperatures <13°C			
	Hours of temperatures >30°C			
	Hours of temperatures >35°C			
	Hours of temperatures between 15°C and 25°C			
	Hours of temperatures between 18°Cand 25°C			
	Hours of temperatures between 18°C and 30°C			
Humidity	Relative humidity	Daily, day and night	average, minimum, maximum	% relative humidity
	Relative humidity of the last 6 days	Daily		
	Hours of relative humidity >95%	Daily, day, night and 6 days	Sum	Hours
	Hours of relative humidity between 70% et 85%			
	Hours of relative humidity between 70% et 95%			
Precipitation	Precipitation	Daily, day and night	Sum	milliliters
	Hours of precipitation	Daily, day, night and 6 days	Sum	Hours