JGR Biogeosciences

METHOD
10.1029/2022JG007342

Special Section:
The Earth in living color: spectroscopic and thermal imaging of the Earth: NASA's Decadal Survey Surface Biology and Geology Designated Observable

# Toward Cloud-Native, Machine Learning Base Detection of Crop Disease With Imaging Spectroscopy

**Gloire Rubambiza[1], Fernando Romero Galvan[2], Ryan Pavlick[3], Hakim Weatherspoon[1], and Kaitlin M. Gold[2]**

[1]Department of Computer Science, Cornell University, Ithaca, NY, USA, [2]Cornell University, Cornell AgriTech, Geneva, NY, USA, [3]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA

**Key Points:**

- Cloud-based plant disease detection system, easily accommodates model improvements and future data sources
- Empower agricultural stakeholders to use hyperspectral data for decision support while preserving stakeholder data privacy
- Outline framework for researchers interested in designing geospatial/remote sensing applications for agricultural stakeholders to follow

**Correspondence to:**

G. Rubambiza and F. Romero Galvan,
gbr26@cornell.edu;
fer36@cornell.edu

**Citation:**

Rubambiza, G., Romero Galvan, F., Pavlick, R., Weatherspoon, H., & Gold, K. M. (2023). Toward cloud-native, machine learning base detection of crop disease with imaging spectroscopy. *Journal of Geophysical Research: Biogeosciences*, *128*, e2022JG007342. https://doi.org/10.1029/2022JG007342

Received 11 DEC 2022
Accepted 26 APR 2023

**Author Contributions:**

**Conceptualization:** Gloire Rubambiza, Fernando Romero Galvan
**Data curation:** Gloire Rubambiza, Fernando Romero Galvan
**Formal analysis:** Gloire Rubambiza, Fernando Romero Galvan
**Funding acquisition:** Hakim Weatherspoon, Kaitlin M. Gold
**Investigation:** Gloire Rubambiza, Fernando Romero Galvan



**Abstract** Developing actionable early detection and warning systems for agricultural stakeholders is crucial to reduce the annual $200B USD losses and environmental impacts associated with crop diseases. Agricultural stakeholders primarily rely on labor-intensive, expensive scouting and molecular testing to detect disease. Spectroscopic imagery (SI) can improve plant disease management by offering decision-makers accurate risk maps derived from Machine Learning (ML) models. However, training and deploying ML requires significant computation and storage capabilities. This challenge will become even greater as global-scale data from the forthcoming Surface Biology & Geology satellite becomes available. This work presents a cloud-hosted architecture to streamline plant disease detection with SI from NASA's AVIRIS-NG platform, using grapevine leafroll-associated virus complex 3 (GLRaV-3) as a model system. Here, we showcase a pipeline for processing SI to produce plant disease detection models and demonstrate that the underlying principles of a cloud-based disease detection system easily accommodate model improvements and shifting data modalities. Our goal is to make the insights derived from SI available to agricultural stakeholders via a platform designed with their needs and values in mind. The key outcome of this work is an innovative, responsive system foundation that can empower agricultural stakeholders to make data-driven plant disease management decisions while serving as a framework for others pursuing use-inspired application development for agriculture to follow that ensures social impact and reproducibility while preserving stakeholder privacy.

**Plain Language Summary** Agricultural decision-makers need reliable access to accurate data to make sustainable crop management choices. This is especially important for decisions related to crop disease management, which can have major financial, environmental, and societal impacts. Forthcoming hyperspectral satellite systems such as Surface Biology & Geology will provide spectroscopic imagery (SI) that can be used in combination with machine learning (ML) for agricultural decision making at the global scale. However, deploying ML models trained on SI requires significant computation and storage resources, limiting non-expert use. Additionally, agricultural stakeholders frequently have reservations and/or restrictions about how data can or cannot be shared with outside entities. Here, we overview a proof-of-concept, cloud system designed with agricultural users in mind that allows researchers to rapidly deploy ML models for plant disease detection using SI from AVIRIS-NG without retaining confidential stakeholder information. We use grapevine leafroll virus-complex 3 (GLRaV-3) in California wine grapes, a virus that causes $3 billion in damages and losses to the US grape industry annually, as a case study. We provide a framework design that outlines how this system is implemented and could be made accessible to both growers and researchers, as well as discuss system limitations and opportunities for future work.

## 1. Introduction

Food security through consistent and scalable agriculture output is the invisible foundation of modern society. However, the globalization and interconnectedness of the agricultural problems we now face demand global, interconnected, and scalable solutions. For example, climate change reduced global farming productivity by 21% over the past 60 years while food demand increased 100% (Ortiz-Bobea et al., 2021). Food demand is anticipated to rise an additional 60%–90% over the next 30 years as the population continues to grow (FAO et al., 2015). Plant disease threatens our ability to scale agricultural output to meet this continuously growing demand. Currently, plant disease destroys an estimated 15%–30% of the global harvest annually, resulting in $220B in losses (Secretariat et al., 2021). Climate change is anticipated to increase the virulence, geographic range, and dispersal capacity of plant pathogens, and consequently, their downstream impacts to human health

**Methodology:** Gloire Rubambiza, Fernando Romero Galvan
**Resources:** Gloire Rubambiza, Fernando Romero Galvan, Ryan Pavlick, Hakim Weatherspoon, Kaitlin M. Gold
**Software:** Gloire Rubambiza, Fernando Romero Galvan
**Supervision:** Ryan Pavlick, Hakim Weatherspoon, Kaitlin M. Gold
**Validation:** Gloire Rubambiza, Fernando Romero Galvan
**Visualization:** Gloire Rubambiza, Fernando Romero Galvan
**Writing – original draft:** Gloire Rubambiza, Fernando Romero Galvan
**Writing – review & editing:** Gloire Rubambiza, Fernando Romero Galvan, Ryan Pavlick, Hakim Weatherspoon, Kaitlin M. Gold

and society (Nnadi & Carter, 2021). Providing agricultural stakeholders with privacy-preserving, easily deployable, and scalable methods to detect early-stage plant disease, the period of time when intervention is both most critical and most likely to succeed can help address this challenge and reduce food insecurity.

Spectroscopic imagery (SI) in the visible to shortwave infrared light range (VSWIR, 400–2,400 nm) can quantify chemistry in soil, rock, and vegetation based on the interaction of light with chemical bonds (Curran, 1989). Plant disease changes how solar radiation interacts with leaves, canopy, and plant energy balance, all of which is captured in SI. This underlying capacity is what enables airborne imaging spectroscopy to non-destructively detect biotic stress in both natural and agroecosystems asymptomatically (Romero Galvan et al., 2023; Sapes et al., 2022; Zarco-Tejada et al., 2018, 2021). Disease detection with SI has benefited greatly from data processing using artificial intelligence (AI), specifically machine learning (ML) (Hruška et al., 2018; Jiménez-Brenes et al., 2019). However, using ML with SI requires powerful compute and storage media to process potentially terabytes (TBs) of data. Imaging spectrometers such as the Airborne Visible/Infrared Imaging Spectrometer Next Generation (AVIRIS-NG) instrument have collected SI over millions of acres of agricultural lands unintentionally during campaigns targeted at other uses. Forthcoming satellite systems such as the European Space Agency's Copernicus Hyperspectral Imaging Mission for the Environment (CHIME; Nieke & Rast, 2018) and NASA's Surface Biology and Geology (SBG); (Schneider et al., 2019), will revolutionize global imaging spectroscopy data availability. Taken as a constellation, these instruments will provide data at actionable intervals without cost, and will, for the first time, democratize the availability of such powerful data products for agricultural use. The AVIRIS-NG archives therefore present an exciting opportunity to test and validate the utility of SI for disease management decision making. However, deploying cutting edge models for agricultural stakeholder use beyond academic investigations requires a flexible infrastructure that can readily access a user's edge devices, such as computers, storage, and networks. Cloud computing is an ideal solution to this challenge.

Cloud computing has revolutionized data storage and processing by providing ubiquitous, convenient, and on-demand network access to potentially unlimited storage and compute devices (Mell & Grance, 2011). In fact, a third of the world's data, compute, and storage is hosted by the cloud (Cohen, 2021; Marr, 2015). Practically, the cloud is a collection of servers dispersed in regional or global data centers to provide low-latency computations (e.g., AI and ML) closer to users in nearly every industry. For instance, the US National Football League (NFL) relies on Amazon Web Services (AWS) to improve the fan experience and player safety (Amazon Web Services, 2022). In shopping and entertainment, the cloud and ML are the backbone of catalog recommendations or movie recommendations streaming services. The most profitable segment of the cloud is on-demand access to servers and general-purpose ML models provided by AWS or similar offerings from competitors such as Azure Cloud, Google Cloud Platform, and IBM Cloud (Google, 2022; IBM, 2022; Microsoft, 2022). In agriculture, the cloud and AI are increasingly employed in yield prediction, soil mapping, land and data management, etc (Analytics, 2022; Awan et al., 2020; Pavón-Pulido et al., 2017). In the context of SI-informed disease management, cloud computing is not only needed to scale current methods, but also to explore the implications of future missions. Agricultural decision-making stands to benefit from ML models trained using powerful cloud computers. However, in contexts where Internet connectivity is sparse, as is often the case in rural communities, it is more practical for models to be trained in the cloud but deployed closer to the data sources. This distributed computing model, which complements cloud computing, is known as edge computing. The edge cloud is a similar compute system organization as the cloud, but rather than providing all computing and storage resources on remote servers, these capabilities are, in part, local and semi-autonomous to accommodate limitations in centralized systems and the intermittency or complete loss of Internet connectivity.

The goal of this work is three-fold. First, we aim to provide a general, cloud-based platform for training and testing SI-informed ML models for non-intrusive and scalable crop disease management. To that end, we showcase new techniques for processing Section 2.4 and refining spectroscopic imagery Section 2.3 in a novel agricultural application: grapevine leafroll associated virus complex 3 (GLRaV-3) detection in wine grape Section 2.1 (Romero Galvan et al., 2023). The second goal is to provide a general platform for researchers to explore the implications of future satellite missions, including SBG and beyond, whose data scales and resolutions are yet to be finalized, for actionable disease detection and decision making. We demonstrate a cloud-based software architecture whose underlying data and application programming interfaces (APIs) are “plug-and-play” to accommodate current known, and future, yet unknown needs Section 2.2. Finally, the final goal of this work is to achieve a positive social impact by providing agricultural stakeholders, including grape growers and industry members, with insights from our work through a platform that is not only easily accessible but also designed with

**Figure 1.** Cabernet Sauvignon grapevine infected with GLRaV-3; one leaf (top of image) not showing the characteristic GLRaV-3 foliar red-blotch symptoms while another leaf (center of image) is showing characteristic red-blotch foliar symptoms of GLRaV-3 infection.

data privacy in mind. To this end, we illustrate the efficacy of ML models developed for GLRaV-3 detection (Figure 6) in a cloud-native environment that does not retain potentially proprietary grower data (e.g., exact field locations) within it. Finally, we discuss methods for reproducibly sharing these models in a manner as required by public funding agencies, while maintaining grower privacy and support Section 2.2.3.

## 2. Materials and Methods

### 2.1. Model Pathosystem, Disease Incidence, and Detection Models

Grapevine leafroll-associated virus complex 3 (GLRaV-3) is an economically important viral disease of wine, juice, table, and raisin grape. GLRaV-3 is estimated to cause up to 3 billion (USD) in economic damage to the U.S. wine and grape industry annually. While the virus is primarily vectored by mealybugs (Pseudococcidae sp.), it can also be spread by many other phloem feeding insects (J. Charles et al., 2006; J. G. Charles et al., 2009; Golino et al., 2002; Pietersen et al., 2013). Vines infected with GLRaV-3 ultimately have reduced lifespan and productivity. Additionally, infection causes uneven berry ripening and disordered berry chemistry, which reduces wine quality. Detecting GLRaV-3 infected grapevine at the symptomatic stage is straight forward: scouts are trained to recognize the foliar symptoms (see Figure 1) and a piece of the grapevine is sent over to a commercial lab to verify viral presence via serological or molecular testing. However, the main challenge growers face is early detection. Infected plants can stay at the asymptomatic stage for up to 1 year, meaning they display no visible sign of disease for humans to identify despite being infectious to nearby grapevines (Almeida et al., 2013; Maree et al., 2013; Naidu et al., 2014). Additionally, unlike red grape varieties (e.g., Cabernet Sauvingnon), white grape varieties (e.g., Chardonnay) do not manifest visible, foliar symptoms that can be identified by scouts (J. Charles et al., 2006; Naidu et al., 2014; Olmos et al., 2016). Existing methods for asymptomatic GLRaV-3 rely on molecular testing which can cost $40–$300 USD per vine depending on the number of viruses that are being tested. Considering that a small vineyard has close to 1,000 vines, and a large vineyard 30,000+, this presents a crucial scaling problem that imaging spectroscopy is ripe to address.

All data worked with here is within the city of Lodi, California, USA illustrated in Figure 2. Here, there are roughly 11,000 acres of vineyards where AVIRIS-NG acquisitions were made. The AVIRIS-NG acquisitions captured vineyards roughly a week after the disease incidence coordinates were recorded. For more detailed description of disease incidence collection, validation, and aggregation, see Romero Galvan (Romero Galvan et al., 2023). In brief, disease incidence data was collected at seven vineyards and covered 7 acres (ac) of red grape variety Aglianico, 204ac of Cabernet Sauvignon, and 57ac of Petite Sirah in the city of Lodi, CA by visual inspection for foliar symptoms (“scouting”) by expert teams. A subset of the data was sent for external validation via molecular testing in September of 2020 and 2021. Molecular testing found the expert scouts to be 100% accurate. When a diseased vine was encountered, scouts recorded the coordinates of disease incidence in local Universal Transverse Mercator geographic coordinate system, Zone 11 (EPSG: 26910). Each disease incidence point is classified according to one of three labels: non-infected (NI), symptomatic (Sy), and asymptomatic (aSy). It is well established that GLRaV-3-infected red variety grapevines (e.g., Cabernet Sauvignon, Malbec) can stay at the asymptomatic stage for up to a year if no visible symptoms appear before a grapevine's winter dormant period (J. Charles et al., 2006; Olmos et al., 2016). Therefore, incidence points identified as visibly infected in 2021 were labeled asymptomatic (aSy), and vines identified as visibly diseased by the scouts in 2020, during the AVIRIS data acquisition, are labeled as symptomatic (Sy). Vines that were identified as Sy in 2020 were removed prior to the 2021 growing season, and therefore were not present at the time of 2021 scouting. Vines that were not identified as visibly infected in either 2020 or 2021 are labeled non-infected (NI). We note that we did not conduct molecular testing to prove that these vines were truly aSy at the time of data collection, as it would

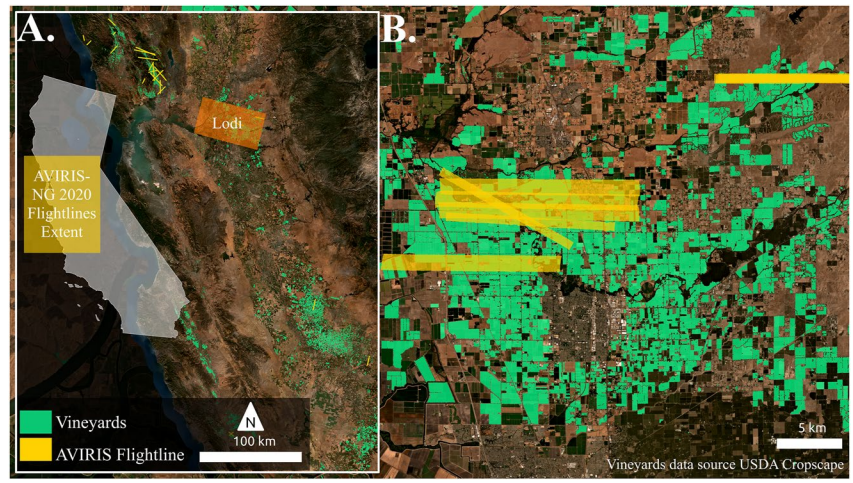


**Figure 2.** Area of interest (AOI) highlighting AVIRIS-NG acquisitions which spatially overlap our ground validation data in the city of Lodi, California. (a) Shows the full geographic extent of all AVIRIS-NG flight lines collected in 2020 that flew over vineyards. (b) Shows a zoom-in to the flight lines that collected SI over vineyards in the AOI.

have been unfeasible given our scope of study and sample testing expense ($40–50 per vine). We emphasize that this assumption is well supported by current understanding of disease biology (Almeida et al., 2013; Maree et al., 2013; Naidu et al., 2014), and the fact that all green foliage was destroyed and removed from the vineyard during mechanical harvest soon after the flight took place. This implies a lower likelihood that vines experienced an opportunity to become infected between the time of the AVIRIS-NG flight and bud break the following season.

Following common convention of training ML models (Gholamy et al., 2018), our Random Forest models (RFm) were trained through a 70/30 training/validation split. The models were trained on spatially-resampled 3-m SI as Romero Galvan found to be optimal for GLRaV-3 detection in grapevine (Romero Galvan et al., 2023) as it was the optimal balance of not losing too much of the disease signal and drowning the noise introduced by the soil background present at each pixel. In total, 268 acres were classified, capturing 80% of the disease incidence in the field.

### 2.2. System Architecture

The system architecture consists of three major components that provide an adaptable pipeline for disease detection. These components are the NASA Cloud, the edge lab at Cornell University, and the publicly accessible cloud, specifically, Azure Cloud (Figure 3). We describe these components and the corresponding steps below.

#### 2.2.1. NASA Cloud

The NASA Cloud serves as the peripheral location of raw data. The raw data is collected through NASA's AVIRIS-NG flight missions. That is, the servers host high-dimensional SI data originating from flights over 875,000 acres of vineyards in California, USA. In **Step 1** the SI is accessed from the AVIRIS-NG data portal through a user-initiated download. These data sets serve two distinct functions. First, they serve as data to train new disease detection models. Second, they function as raw data for filtering during field-specific inference requests.

#### 2.2.2. Edge Lab

The edge lab serves as a central location for data pre-processing. Before the NASA Cloud data can yield useful insights, it requires pre-processing to account for noise in the data collection process. The pre-processing steps must include those highlighted in Section 2.4: BRDF, topographic, vector normalization, etc. all steps are illustrated in Figure 4. In **Step 2**, the pre-processed training data is uploaded to the cloud for model training. In addition to data pre-processing, the edge lab serves as training ground for the GLRaV-3 model training/testing. Note that, in **Step 3**, the models can be trained in the cloud or at the edge lab. On one hand, in **Step 4**, the training relies on pre-processed data and more compute power in the cloud. On the other hand, the edge lab (or "local")

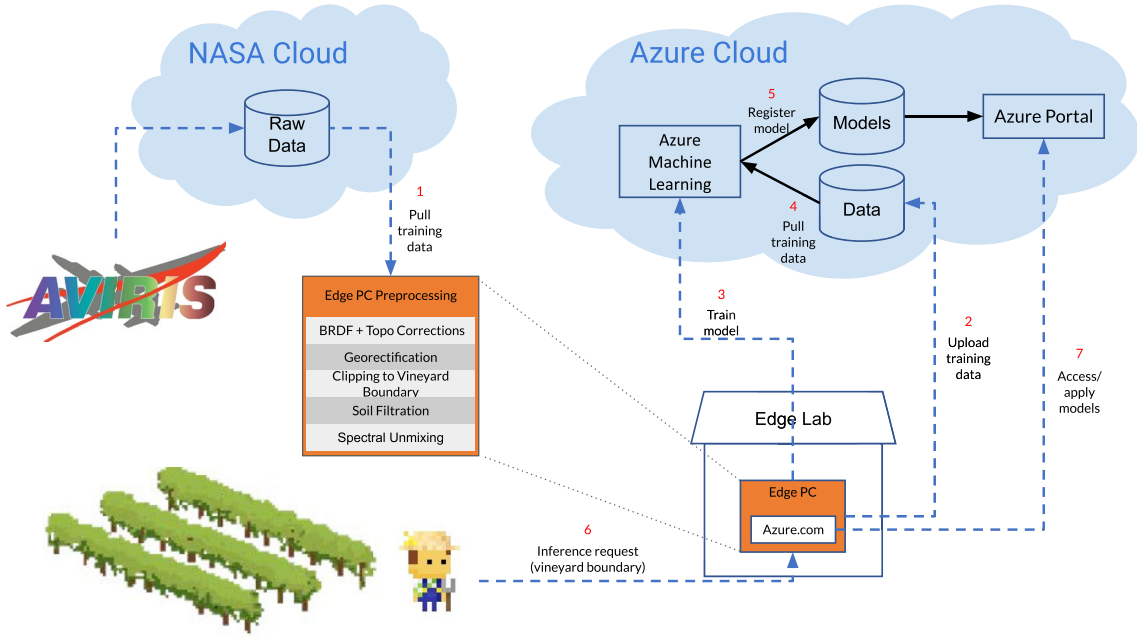


**Figure 3.** Flowchart of the system architecture for early GLRaV-3 disease detection.

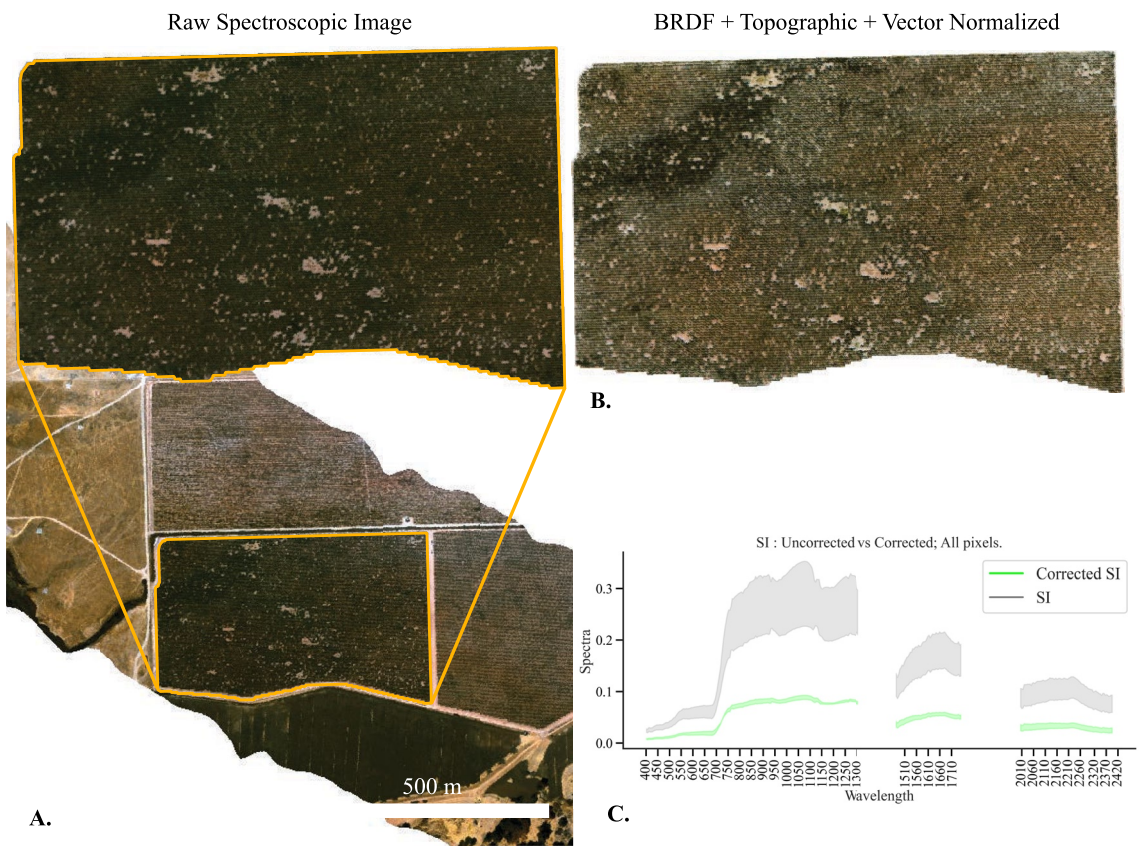


**Figure 4.** (a) AVIRIS-NG Flightline with inset of example vineyard as seen by AVIRIS-NG. (b) Imagery with BRDF, topographic, and vector normalization applied. (c) Spectroscopic graph of reflectance values before and after.

training and testing follows from three practical considerations. First, it affords researchers as much flexibility as possible in setting up environments in terms of software dependencies. Second, to minimize cloud consumption costs, disease inference requests can be iteratively tested locally before being deployed to the cloud. Lastly, though rare in research labs, the connectivity to the public cloud for data and model access may be limited, and the local setup may intermittently serve inference requests during such outages. The training at the edge lab yields GLRaV-3 models that are managed by the model management pipeline described next.

#### 2.2.3. Model Management Pipeline

In **Step 5**, to store the models in the cloud, they must be serialized (a process also known as "pickling"). Serialization is the process of turning objects in a computer program into bits that can be sent on a network like the Internet and/or stored in persistent storage like a hard disk drive (HDD) or solid-state drive (SSD). The models essentially represent mathematical functions that represent the most salient components in the disease detection process. Therefore, the serialization process captures these functions as (Python) objects that can be uploaded from one site and downloaded at another site with the same representation. In this context, the GLRaV-3 models can be uploaded/downloaded locally or stored in the cloud for retrieval during disease inference requests (**Step 6**).

The model upload process in Step 5 specifies the location of the model on the local computer, the name for the model, any appropriate tags, and the Azure Machine Learning (ML) workspace where the model is to be uploaded. Azure ML is a platform-as-a-service (PaaS) offering to enable data processing in the Azure Cloud, and a workspace is a logical partition of computing and storage servers for a single project (Microsoft, 2021). Therefore, the name of the model must be unique within a workspace. The tags are useful to filter models during the download process. Within the workspace, models are indexed by names and versions. The versions are automatically created by the Azure ML workspace when an existing model name is registered more than once. In other words, all models are stored under a single location, but it is possible to specify which version to download and apply for inference requests.

The download process can be triggered programmatically or manually. The programmatic download can be triggered from any (Python) script with the correct credentials for the workspace where the model is stored. The programmatic trigger is ideal for downloading the model at the network edge where connectivity to the cloud is limited in order to field inference requests even during network outages. As demonstrated in **Step 7**, the manual download is possible from any networked computer with access credentials for the Azure ML workspace through the Azure portal. The Azure portal is a user-friendly interface for provisioning and managing resources in the Azure Cloud. The manual download is a convenient way to access the model parameters and potentially check the size before targeting a model deployment location or device. This is important for inference requests that may be run on resource-constrained devices such as Raspberry Pis (Raspberry Pi, 2022).

### 2.3. Spectroscopic Imagery

Spectroscopic Imagery (SI) from NASA Jet Propulsion Laboratory's AVIRIS-NG airborne instrument was used to train these models. Specifically, the SI image IDs were: ang20200918t210249, ang20200918t205737, ang20200918t212656, ang20200918t213801, ang20200918t213229. The SI imagery was collected on 18 September 2020 between 1:00 p.m. and 3:00 p.m. Pacific (local) time. All listed SI are collected at the 1-m spatial resolution, collected spectral-channels at 5-nanometers (nm) between 380 and 25,000 nm totaling 425 channels (Chapman et al., 2019; Thompson et al., 2018). All AVIRIS-NG imagery used in this study is publicly available reflectance data that can be downloaded from the AVIRIS-NG data portal https://aviris.jpl.nasa.gov/dataportal/. All AVIRIS-NG SI were atmospherically corrected using the ATmospheric REMoval (ATREM) algorithm (Gao & Goetz, 1990). Additionally, water absorption feature wavelengths are excluded due to noise caused by atmospheric affects and water absorption featured found at these sections of the electromagnetic spectrum. To be specific, the spectroscopic ranges excluded from the training data: 380–400 nm, 1,310–1,470 nm, 1,750–2,000 nm, and 2,400–2,600 nm. All pixels outside the extent of the vineyard are likewise excluded by masking the imagery using the Python package Rasterio (Gillies et al., 2013). To exclude non-vine spectra from the imagery (e.g., soil, shadows, trees), a vine mask was generated using the spectral unmixing residuals outlined in Sousa et al. (2022). Both the bidirectional reflectance distribution function (BRDF) and topographic corrections (Queally et al., 2022) scripts from the HyTools package (Chlus et al., 2022) were applied to all of the SI using default parameters. BRDF and topographic corrections are common techniques in remote sensing to

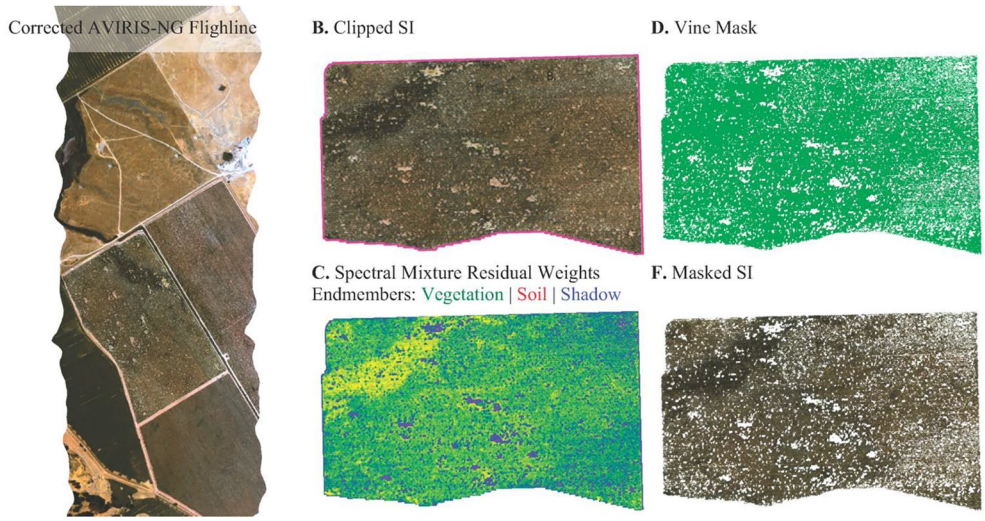


**Figure 5.** (a) Full AVIRIS-NG corrected SI. (b) SI clipped to the boundary of the vineyard. (c) SMR endmember weights. (d) Vine mask generated from the SMR weights. (f) Resulting SI image from the pipeline which excludes non-vegetation pixels.

address sources of error and noise. BRDF accounts for differences in how surfaces reflect light in different directions, which can cause distortions in the image. Topographic corrections account for the effects of terrain on the reflection of light. A subset of the AVIRIS-NG imagery required further co-registration to improve the spatial alignment of the SI with the disease incidence scouting geographic coordinates. The National Agriculture Imagery Program (NAIP; USDA) imagery collected within a week of our AVIRIS-NG imagery was used as a reference to improve georeferencing and co-registration. Ground control points (GCPs) were generated by passing a target SI and reference NAIP image to the open-source Python library Automated and Robust Open-Source Image co-registration Software (AROSICS) (Scheffler et al., 2017). Lastly, the n-dimensional spectroscopic vectors were normalized to reduce the effects of varying illumination conditions that may be present in the field. The reason for applying vector normalization is to make uniform the brightness of each band of the SI. This results in a normalized SI where the values of each band are only influence by the spectral characteristics of the surface rather than illumination conditions. Vector normalization was done using the normalization function from the Python library Numpy to calculate the magnitude or "length" of each vector, each vector is then divided by its length resulting in a normalized vector: $|\mathbf{v}|$ where $\mathbf{v}$ is the vector and $|\mathbf{v}|$ is the magnitude, which results in a unit vector in the direction of $\mathbf{v}$. These corrections are illustrated in Figure 4, note the comparison of illumination in the raw SI (a) versus the corrected image (b), as well as the values of the spectra after normalization (c).

### 2.4. Processing SI Data

Our pipeline takes AVIRIS-NG Imagery with the corrections in Figure 4b already applied. This automatic pipeline takes AVIRIS-NG Imagery and outputs a classified geospatial raster. The pipeline steps are illustrated in Figures 5a–5f. The first step masks the AVIRIS-NG SI to the extent of the vineyard boundaries. Spectral residual weights are then used to generate a vegetation mask Figures 5a–5d. The resulting mask is used to remove pixels where the spectral signal of vegetation is at least 50% of the total spectral contribution Figure 5. The resulting image is now ready for it to be classified using the RFm generated in Section 2.1.

## 3. Results

The results of this study are threefold. Frist, we showcase the deployment of a cloud-hoster GLRaV-3 detection system (Figure 3) capable of training and applying GLRaV-3 detection RF models to AVIRIS-NG spectroscopic imagery in the edge cloud (e.g., the farm) and public Azure cloud (Section 3.1). Second, we demonstrate the effectiveness of RF models trained following the methodology outlined (Romero Galvan et al., 2023) and deployed to the system (Section 3.2). Lastly, we present a release of executeable python scripts for the training and application of these models to an AVIRIS-NG image (Galvan & Rubambiza, 2023).

**Table 1**
*Grapevine Disease Detection: Model Training Runtime and Accuracy Using Various Configurations*

| Training | Data location | Data size | Storage (GB) | Runtime (s) | Accuracy (%) |
|---|---|---|---|---|---|
| Edge | Local | 4MB | 256 | 27.1 | 84 |
| Edge | Cloud | – | 256 | 35.6 | 86 |
| Azure ML | Cloud | – | 100 | 86.5 | 86 |

### 3.1. GLRaV-3 Detection Across Clouds

In building the cloud-based early disease detection system, we aim to demonstrate an end-to-end platform. Therefore, we showcase trade-offs between disease detection in the public cloud and the edge cloud. That is, we compare the resulting latency and model accuracy in executing disease inference request with models/data deployed in the public cloud and edge clouds. Specifically, we set up four experimental training/testing conditions: (a) cloud-based model and cloud-based SI data (control setup), (b) local model and local SI data, (c) local model and cloud-based SI data. The edge cloud setup (i.e., conditions a and b) are desirable during periods of intermittent Internet connectivity using local data and/or models. On the other hand, the public cloud can scale to meet numerous disease inference requests at once. Thus, we set up an experimental environment using virtual machines (VM) in the public cloud and a PC representing the edge cloud.

Table 1 demonstrates that model accuracy is stable and decoupled from the training/inference location. Further, executing disease detection requests with local data and models incurs the least amount of latency (27.1 s). In executing a request at the edge with SI data in the cloud compared to local SI data, we observed a 31% overhead (i.e., 27.1 vs. 35.6 s) which is likely due to a "warm-up" period of downloading training data from the public cloud. In other words, subsequent requests using a local model and public cloud-based data were relatively similar to loading data from the local disk. The results demonstrate different data/model placements available to agricultural stakeholders with different levels of robustness.

### 3.2. GLRaV-3 Detection Maps

The final output of the model management pipeline are the classified rasters for which we provide an illustrative example in Figure 6. These rasters contain the relevant disease incidence stage Figure 6a and the model's prediction confidence Figure 6b for each 3-m pixel. The pipeline produces a total of six rasters, consisting of three pairs. The first pair includes rasters for non-infected vine locations and infected vines at either the symptomatic or asymptomatic stage, each with a class probability raster to show the model's confidence in the label. The second pair contains a raster for non-infected and symptomatic vines, also with a corresponding class probability raster. The third pair has a raster for non-infected and asymptomatic vine locations, along with its class probability raster. The rasters are georeferenced, meaning that each raster is aligned with the geographic coordinate system of the input AVIRIS-NG image. As a result, these rasters are compatible with Geographic Information Systems (GIS) such as ArcMap or QGIS, and can be further analyzed through these tools.

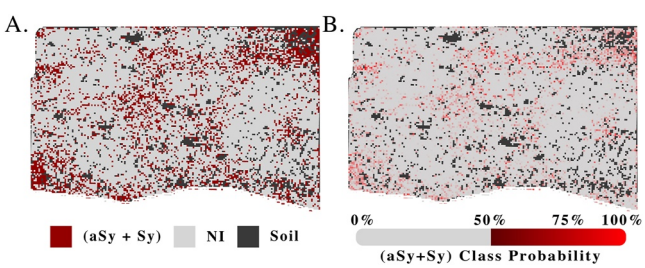


**Figure 6.** Example of the output prediction rasters from the cloud system architecture. These rasters illustrate the GLRaV-3 incidence predictions made using an AVIRIS-NG image clipped to the boundaries of a commercial vineyard in Lodi, California. (a) The output rasters for disease incidence, in red, are the GLRaV-3 incidence classifications. In gray, are the non-infected predicted vines. In black are soil pixels that were masked out. (b) The output probability rasters of the (aSy + Sy) class, where pixels that reach a minimum of 50% confidence score are highlighted from dark-red (50%) to bright-red (100%).

## 4. Discussion

Successful detection of GLRaV-3 at the asymptomatic stage provides grape growers up to a year of warning to act on removing the infected vine before any further spread could take place. Our methodology and results showcase a robust system with significant implications for decision support not only for crop disease detection but also for future data sources (e.g., SBG) and use cases targeted to improve our understanding of disease incidence processes at a global scale. First, the proposed system architecture is general enough to enable plug-and-play of different and complementary models. For example, this architecture can easily be reused to train and test grape variety classification models, which in turn we can use to better refine our methodologies for disease detection across grape varieties. This adaptability may prove important for both future application and model improvement, given different management requirements and innate disease resistance across grape varieties which may affect the spectroscopic signal, as illustrated in Figure 7. This figure illustrates the unique spectral signatures that exist between grape varieties Cabernet Sauvignon and Chardonnay. Two vineyards with AVIRIS-NG SI were used to generate this graph. Both vineyards are right next to each

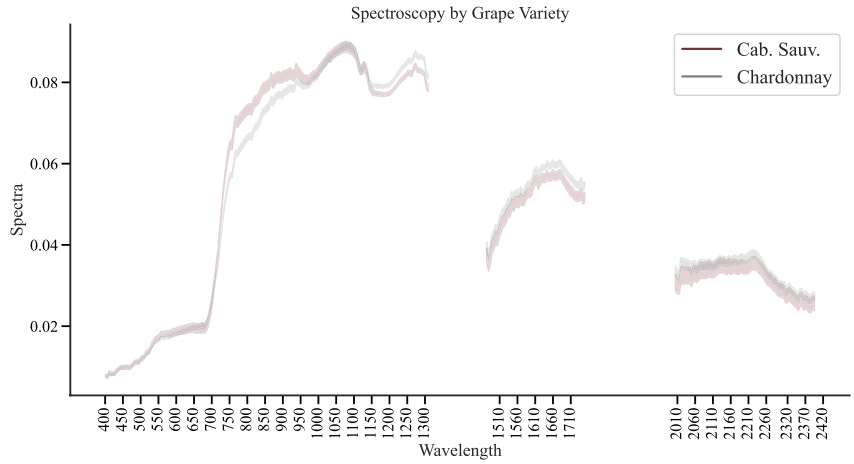


**Figure 7.** Spectroscopic signal for red-grape variety Cabernet Sauvignon and white-grape variety Chardonnay. Here we plot the spectra of two vineyards by variety and show the maximum and minimums at each wavelength.

other and are managed by the same vineyard owner ruling out potential sources for these spectral differences. Thus, suggesting that developing grape variety classification models are possible and may indeed help disease detection models improve their accuracy. Second, beyond model variety, we designed the software architecture to be plug-and-play to different data sources. Currently, the system provides GLRaV-3 decision support at 3 m resolution with SI derived from AVIRIS-NG. As we look to future satellite missions, such as SBG and CHIME, our system can be leveraged to explore disease detection efficacy at these systems' anticipated resolutions (30 m) years ahead of their launches.

In our context, the public cloud provides storage and computation spaces for the disease detection data and models (both of which are refined by the edge lab) (Mell & Grance, 2011). It is important to note that although we use the Azure Cloud, the "Google Cloud," "Amazon Cloud," and "IBM Cloud," could have been used as well. Our use of cloud computing plays a crucial role in fulfilling our architecture goals to have social impact and research reproducibility while preserving user privacy. Social impact is achieved by demonstrating a holistic, simple architecture that agricultural stakeholders can access as a service in the decision-making process via refined models. The models can be deployed as web endpoints that can be called upon to perform field data inference requests on edge devices. Our architecture also provides a platform that contributes to reproducible research while addressing user data privacy concerns by leaving all potentially proprietary data on the user's edge device. Through the Azure ML platform that the architecture wraps around, researchers can access and share disease detection models within both their community of peers and stakeholders (Microsoft, 2021). As discussed in Section 2.2.3, Azure ML enables systematic registration, tracking, and tagging of models by researchers in the course of the experimentation process that is at the core of the scientific method. From iterative and tagged models, researchers can confidently make general and reproducible knowledge claims. While this architecture makes strides toward reproducibility, streamlined sharing of models (e.g., access permissions, appropriate application programming interfaces, or APIs) are still needed to enable true access and sharing not only across institutions but also across cloud providers. This challenge is not addressed in this work, and it remains an open area of exploration for our future work in this vein (see Section 4.1).

### 4.1. Limitations

Co-occurring biotic and abiotic stresses can confuse SI-informed GLRaV-3 detection (Romero Galvan et al., 2023). It is not uncommon for a vine, or multiple neighboring vines, to be infected with multiple pathogens (e.g., trunk disease), and/or experience co- occurring abiotic stress (e.g., drought). Further work is required to gauge how common abiotic stressors in Californian vineyards (e.g., water stress, insect damage, or sunburn) cause misclassifications. However, as long as the model correctly identifies a plant as abnormal, this information is still of great value to agricultural stakeholders because it can be used to strategically deploy scouts and field management teams to areas most in need of attention. Thus the potential service outlined here is still valuable to the grape grower community, despite this open area for improvement. Another limitation is AVIRIS-NG acquisition

availability. While the majority of grapevine in the United States is in California, the home of AVIRIS-NG and thus the state with the most historical coverage, there are growers in other regions including New York, which is the 3rd most important grape-growing state in the US. AVIRIS-NG is an airborne platform that requires appropriate flying conditions for an acquisition to happen, and thus can be both expensive and logistically challenging to deploy. However, spaceborne imaging spectrometers such as SBG, Soil Mapper, and CHIME will soon provide global SI. With the help of this proposed framework, we open the door for global-scale GLRaV-3 monitoring beyond the currently outlined geographic area of study.

In this work, we have demonstrated the design, prototype, and inference results of an Azure Cloud-based disease detection system. The next logical step is to streamline the system toward a more accessible user experience for agricultural stakeholders and researchers. However, we face two limitations. First, although Azure provides the capability to deploy the models as web-based endpoints (e.g., issuing web requests with shape files and getting back classified rasters of disease presence), the disease detection system is still tied to a single cloud provider. To avoid technical compatibility issues inherent in vendor lock-in (e.g., transferring data across private and public clouds), the system must evolve to work across clouds so that the data and models are not siloed in a single cloud. Second, the system's effectiveness has been shown in a few fields. In practice, growers operate thousands of fields potentially spread across multiple farms. One area that remains unexplored is managing diverse models across farms. For instance, growers can operate water-stress models at some farms while deploying GLRaV-3 models on other fields. Technologically, this necessitates scalable techniques for managing the model and data lifecycle across farms and is the subject of ongoing future work.

## 5. Conclusions

Here, we describe and deploy an adaptable cloud-based system for detecting an economically important crop disease, GLRaV-3 in grapevine as a case study, with airborne imaging spectroscopy data. The proposed architecture only requires SI from an airborne or spaceborne source and a shapefile from the user specifying location coordinates and boundaries to run. The goal of this system is to empower agricultural stakeholders to make well-informed, data-driven decisions by granting them access to the latest advances in SI-ML disease detection. Our work improves crop disease decision-making while serving as a guide for others interested in developing accessible, use-inspired, remote sensing applications for agricultural stakeholders who stand to benefit from publicly funded research. Besides the exploration of different models and data modalities, the system could fundamentally change the way we approach vineyard management operations. Infected vines, be they symptomatic or asymptomatic, still produce grapes, albeit of lower quality for the wine making process, which chips away at profit across the value chain. Typically, asymptomatic vines are hard to detect, which ultimately affects grape and wine quality. The system presented in this work can provide a valuable predictor/indicator of potential yield, as well as which vines are spectroscopic anomalies, signaling stress, disease, or any other damage. This predictive analytics service is useful in anticipating the chemical supplementation that may be necessary for the upcoming harvest and fermentation of lesser-quality grapes. From a financial perspective, molecular testing, the most reliable method of asymptomatic detection, does not scale to the sizes required. However, armed with a disease detection system, growers can strategically deploy their high-accuracy ground resources, such as expert scouts or molecular testing, more efficiently because the system predicts with high accuracy vineyard regions most likely to be infected. In this manner, the system complements the farm ecosystem, empowering the users, and not seeking to replace their valuable expertise.

## Data Availability Statement

All spectroscopic imagery used for this study are publicly available without restrictions from the AVIRIS-NG Data portal: https://aviris.jpl.nasa.gov/dataportal/. v0.0.1 of the GLRaV3Detection Pipeline used for hosting the GLRaV-3 detection models and applying scripts for applying the GLRaV-3 detection models on the AVIRIS-NG imagery is preserved at https://doi.org/10.5281/zenodo.7826842, available via and developed openly at https://zenodo.org/badge/latestdoi/573130843.

**Acknowledgments**
This work was funded in part by NASA FINESST (Grant 80NSSC21K1605) awarded to Romero Galvan (FI) and Gold (PI), NASA Jet Propulsion Laboratory Strategic University Research Partnership Fund, NSF NRT Digital Plant Sciences training Grant awarded to Cornell University (Grant 1922551), the NSF STC Center for Research on Programmable Plant Systems (CROPPS; Grant 2019674), and the NSF CHS Grant understanding and improving the social impact of high-bandwidth farm networking infrastructure (Grant 1955125). We acknowledge with gratitude Dr. Michael Scanlon, project leader of the Cornell University NSF NRT Digital Plant Sciences program that introduced and inspired Rubambiza and Romero Galvan to collaborate on this co-first authored paper. Additionally, we thank Dr. Michael Eastwood and the myriad members of the NASA JPL AVIRIS-NG team who, along with the NASA Biodiversity and Ecological Conservation program office, are responsible for initiating and executing the successful air campaign that provided data for this publication. Importantly, we would like to thank our industry collaborators, including Charlie Starr, Stephanie Bolton, Mimar Alsina, and Nick Dokoozlian, for their invaluable efforts, feedback, and collaboration that supported this project. A portion of this research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration (80NM0018D0004).

## References

Almeida, R. P. P., Daane, K. M., Bell, V. A., Blaisdell, G. K., Cooper, M. L., Herrbach, E., & Pietersen, G. (2013). Ecology and management of grapevine leafroll disease. *Frontiers in Microbiology*, *4*. https://doi.org/10.3389/fmicb.2013.00094

Amazon Web Services, I. (2022). *NFL next gen stats powered by AWS—Stat that*. Amazon Web Services, Inc. Retrieved from https://aws.amazon.com/sports/nfl/

Analytics, A. (2022). *Ag-analytics: Discover agricultural data*. Analytics.ag. Retrieved from https://www.analytics.ag

Awan, K. A., Ud Din, I., Almogren, A., & Almajed, H. (2020). AgriTrust—A trust management approach for smart agriculture in cloud-based internet of agriculture things. *Sensors*, *20*(21), 6174. https://doi.org/10.3390/s20216174

Chapman, J. W., Thompson, D. R., Helmlinger, M. C., Bue, B. D., Green, R. O., Eastwood, M. L., et al. (2019). Spectral and radiometric calibration of the next generation airborne visible infrared spectrometer (AVIRIS-NG). *Remote Sensing*, *11*(18), 2129. https://doi.org/10.3390/rs11182129

Charles, J. G., Cohen, D., Walker, J. T. S., Forgie, S. A., Bell, V. A., & Breen, K. C. (2006). A review of the ecology of grapevine leafroll associated virus type 3 (GLRaV3). *New Zealand Plant Protection*, *59*, 330–337. https://doi.org/10.30843/nzpp.2006.59.4590

Charles, J. G., Froud, K. J., van den Brink, R., & Allan, D. J. (2009). Mealybugs and the spread of grapevine leafroll-associated virus 3 (GLRaV-3) in a New Zealand vineyard. *Australasian Plant Pathology*, *38*(6), 576. https://doi.org/10.1071/AP09042

Chlus, A., Winstonolson, & Greenberg, E., & Oldmanye007. (2022). *EnSpec/hytools: 1.4.0*. Zenodo. https://doi.org/10.5281/ZENODO.5997755

Cohen, J. (2021). *4 companies control 67% of the world's cloud infrastructure*. PCMAG. Retrieved from https://www.pcmag.com/news/four-companies-control-67-of-the-worlds-cloud-infrastructure

Curran, P. J. (1989). Remote sensing of foliar chemistry. *Remote Sensing of Environment*, *30*(3), 271–278. https://doi.org/10.1016/0034-4257(89)90069-2

FAO, IFAD, & WFP. (2015). The state of food insecurity in the world 2015. *Meeting the 2015 international hunger targets: Taking stock of uneven progress*. Rome: Food and Agriculture Organization of the United Nations. Retrieved from https://www.fao.org/3/a-i4646e.pdf

Galvan, F. E. R., & Rubambiza, G. (2023). *GoldLab-GrapeSpec/GLRaV3Detection: GLRaV-3 detection with AVIRIS-NG imagery pipeline*. Zenodo. https://doi.org/10.5281/zenodo.7826842

Gao, B.-C., & Goetz, A. F. H. (1990). Column atmospheric water vapor and vegetation liquid water retrievals from Airborne Imaging Spectrometer data. *Journal of Geophysical Research*, *95*(D4), 3549. https://doi.org/10.1029/JD095iD04p03549

Gholamy, A., Kreinovich, V., & Kosheleva, O. (2018). Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation.

Gillies, S., & others (2013). Rasterio: Geospatial raster I/O for python programmers. Mapbox. Retrieved from https://github.com/rasterio/rasterio

Golino, D. A., Sim, S. T., Gill, R., & Rowhani, A. (2002). California mealybugs can spread grapevine leafroll disease. *California Agriculture*, *56*(6), 196–201. https://doi.org/10.3733/ca.v056n06p196

Google, I. (2022). *Cloud computing services*. Google Cloud. Retrieved from https://cloud.google.com/

Hruška, J., Adão, T., Pádua, L., Marques, P., Peres, E., Sousa, A., et al. (2018). Deep learning-based methodological approach for vineyard early disease detection using hyperspectral data. In *IGARSS 2018—2018 IEEE international geoscience and remote sensing symposium* (pp. 9063–9066). https://doi.org/10.1109/IGARSS.2018.8519136

IBM. (2022). IBM cloud. Retrieved from https://www.ibm.com/cloud

Jiménez-Brenes, F. M., López-Granados, F., Torres-Sánchez, J., Peña, J. M., Ramírez, P., Castillejo-González, I. L., & de Castro, A. I. (2019). Automatic UAV-based detection of *Cynodon dactylon* for site-specific vineyard management. *PLoS One*, *14*(6), e0218132. https://doi.org/10.1371/journal.pone.0218132

Maree, H. J., Almeida, R. P. P., Bester, R., Chooi, K. M., Cohen, D., Dolja, V. V., et al. (2013). Grapevine leafroll-associated virus 3. *Frontiers in Microbiology*, *4*. https://doi.org/10.3389/fmicb.2013.00082

Marr, B. (2015). *Big data: 20 mind-boggling facts everyone must read*. Forbes. Retrieved from https://www.forbes.com/sites/bernardmarr/2015/09/30/big-data-20-mind-boggling-facts-everyone-must-read/

Mell, P., & Grance, T. (2011). *The NIST definition of cloud computing*. National Institute of Standards and Technology.

Microsoft (2021). What is an azure machine learning workspace. Retrieved from https://www.twilio.com/docs/sms/quickstart/pytho://docs.microsoft.com/en-us/azure/machine-learning/concept-workspace

Microsoft (2022). What is azure—Microsoft cloud services|Microsoft azure. Retrieved from https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-is-azure/

Naidu, R., Rowhani, A., Fuchs, M., Golino, D., & Martelli, G. P. (2014). Grapevine leafroll: A complex viral disease affecting a high-value fruit crop. *Plant Disease*, *98*(9), 1172–1185. https://doi.org/10.1094/PDIS-08-13-0880-FE

Nieke, J., & Rast, M. (2018). Towards the Copernicus hyperspectral imaging mission for the environment (CHIME). In *IGARSS 2018—2018 IEEE international geoscience and remote sensing symposium* (pp. 157–159). IEEE. https://doi.org/10.1109/IGARSS.2018.8518384

Nnadi, N. E., & Carter, D. A. (2021). Climate change and the emergence of fungal pathogens. *PLoS Pathogens*, *17*(4), e1009503. https://doi.org/10.1371/journal.ppat.1009503

Olmos, A., Bertolini, E., Ruiz-García, A. B., Martínez, C., Peiró, R., & Vidal, E. (2016). Modeling the accuracy of three detection methods of *Grapevine* leafroll-associated virus 3 during the dormant period using a Bayesian approach. *Phytopathology®*, *106*(5), 510–518. https://doi.org/10.1094/PHYTO-10-15-0246-R

Ortiz-Bobea, A., Ault, T. R., Carrillo, C. M., Chambers, R. G., & Lobell, D. B. (2021). Anthropogenic climate change has slowed global agricultural productivity growth. *Nature Climate Change*, *11*(4), 306–312. https://doi.org/10.1038/s41558-021-01000-1

Pavón-Pulido, N., López-Riquelme, J. A., Torres, R., Morais, R., & Pastor, J. A. (2017). New trends in precision agriculture: A novel cloud-based system for enabling data storage and agricultural task planning and automation. *Precision Agriculture*, *18*(6), 1038–1068. https://doi.org/10.1007/s11119-017-9532-7

Pietersen, G., Spreeth, N., Oosthuizen, T., van Rensburg, A., van Rensburg, M., Lottering, D., et al. (2013). Control of grapevine leafroll disease spread at a commercial wine estate in South Africa: A case study. *American Journal of Enology and Viticulture*, *64*(2), 296–305. https://doi.org/10.5344/ajev.2013.12089

Queally, N., Ye, Z., Zheng, T., Chlus, A., Schneider, F., Pavlick, R. P., & Townsend, P. A. (2022). FlexBRDF: A flexible BRDF correction for grouped processing of airborne imaging spectroscopy flightlines. *Journal of Geophysical Research: Biogeosciences*, *127*(1), e2021JG006622. https://doi.org/10.1029/2021JG006622

Raspberry Pi, L. (2022). Buy a raspberry Pi 4 model B. Raspberry Pi. Retrieved from https://www.raspberrypi.com/products/raspberry-pi-4-model-b/

Romero Galvan, F., Pavlick, R., Trolley, G. R., Aggarwal, S., Sousa, D., Starr, C., et al. (2023). Scalable early detection of grapevine virus infection with airborne imaging spectroscopy. *Phytopathology®, PHYTO-01-23-0030-R*. https://doi.org/10.1094/PHYTO-01-23-0030-R

Sapes, G., Lapadat, C., Schweiger, A. K., Juzwik, J., Montgomery, R., Gholizadeh, H., et al. (2022). Canopy spectral reflectance detects oak wilt at the landscape scale using phylogenetic discrimination. *Remote Sensing of Environment*, *273*, 112961. https://doi.org/10.1016/j.rse.2022.112961

Scheffler, D., Hollstein, A., Diedrich, H., Segl, K., & Hostert, P. (2017). AROSICS: An automated and robust open-source image co-registration software for multi-sensor satellite data. *Remote Sensing*, *9*(7), 676. https://doi.org/10.3390/rs9070676

Schneider, F., Ferraz, A., & Schimel, D. (2019). Watching Earth's interconnected systems at work. *Eos*, *100*. https://doi.org/10.1029/2019EO136205

Secretariat, I. G., Gullino, M. L., Albajes, R., Al-Jboory, I., Angelotti, F., Chakraborty, S., et al. (2021). *Scientific review of the impact of climate change on plant pests: FAO on behalf of the IPPC secretariat*. IPCC. Retrieved from https://opus.lib.uts.edu.au/bitstream/10453/149380/2/Full%20reportcb4769en.pdf

Sousa, D., Brodrick, P., Cawse-Nicholson, K., Fisher, J. B., Pavlick, R., Small, C., & Thompson, D. R. (2022). The spectral mixture residual: A source of low-variance information to enhance the explainability and accuracy of surface biology and geology retrievals. *Journal of Geophysical Research: Biogeosciences*, *127*(2), e2021JG006672. https://doi.org/10.1029/2021JG006672

Thompson, D. R., Natraj, V., Green, R. O., Helmlinger, M. C., Gao, B.-C., & Eastwood, M. L. (2018). Optimal estimation for imaging spectrometer atmospheric correction. *Remote Sensing of Environment*, *216*, 355–373. https://doi.org/10.1016/j.rse.2018.07.003

Zarco-Tejada, P. J., Camino, C., Beck, P. S. A., Calderon, R., Hornero, A., Hernández-Clemente, R., et al. (2018). Previsual symptoms of Xylella fastidiosa infection revealed in spectral plant-trait alterations. *Nature Plants*, *4*(7), 432–439. https://doi.org/10.1038/s41477-018-0189-7

Zarco-Tejada, P. J., Poblete, T., Camino, C., Gonzalez-Dugo, V., Calderon, R., Hornero, A., et al. (2021). Divergent abiotic spectral pathways unravel pathogen stress signals across species. *Nature Communications*, *12*(1), 6088. https://doi.org/10.1038/s41467-021-26335-3