Forecasting disease spread to reduce crop losses

Akhil Jayant Patil, Shenaha Sivakumar and Raphael Witt
University of Münster (WWU)
Münster, Germany
akhil.patil@uni-muenster.de, shenaha.sivakumar@uni-muenster.de, raphael.witt@uni-muenster.de

Abstract

There are many crop diseases and pests that spread quickly and destroy a significant amount of crops per year in Europe. These outbreaks happen only when the environmental conditions favor these parasites. With the help of satellite imagery, machine learning and forecasting models we intend to reduce crop loss. By applying a machine learning model on satellite images we want to identify patterns in environmental conditions (i.e. temperature, rainfall, humidity, etc.) that lead to the rise or the spreading of such airborne or waterborne diseases and insect plagues. This approach will help forecasting disease spreading and by using remote sensing we plan to identify and track such areas. In our research we did not find any practically implemented tools to predict the spread of crop damaging diseases. The final model will be used by the agricultural department of the governments to take preventive measures in the locations that will be most prone to the infection in coming days as well as by farmers who want to identify problematic areas on their fields. Our model can serve as starting point for bringing this idea to reality. It will help reducing the huge loss of crops in Europe which in turn will benefit the agricultural industry, food production and ultimately European economy.

Keywords: crop disease spread, machine learning, weather forecast, satellite image processing

Introduction

The global agricultural production is not sufficient to meet the demands of exploding populations. Adding more concern to it, around 20-40% of agricultural production suffers damage because of multiple reasons like diseases, animals, etc. [16] and more than 10% loss is due to diseases alone [6]. Crop loss has been linked directly to economic loss on multiple occasions. It can indicate that when there is shortage of agricultural produce in a nation, then they become dependent on imported food, creating further economic deficits. Our idea will help eradicating the wide spread of disease and help the administratives to take precautionary measures in the initially stage itself.

In the paper below we share details of what precise problem we want to address. This is followed by related work done in this domain and projects that try to deal with similar issues. Later, the architecture along with its components and the processing are mentioned. Then we talk about the challenges that we think are likely to occur when this tool is developed and the possible future work in relation to the conception points discussed here. We end the report with a conclusion we had after working on developing this idea within the last months.

Our idea / Focus area

For our research we narrowed this mammoth problem to some extent by focusing initially only on a widely produced crop in Europe and the major disease that it falls victim to. Considering the high production rate, overall consumption and economic importance, we decided to consider wheat crop as our primary subject. The wheat crop production is infected by various types of fungi, animal pest, weeds, pathogens, viruses etc. Diseases like stem rust outbreaks in a big area and spreads out in no time infecting wider adjoining area, thus affecting the growth and production rate of wheat. The stem rust fungus has caused tremendous amount of damage to wheat crop across Europe in past few years [13]. European MARS Crop Yield Forecasting System monitors crop yield and production [22]. But the focus of our idea is to predict the pattern of the spreading or transmission of the diseases from the point of its origin. Thus, it will be easier for the agricultural department to take immediate action or preventive measures in averting major damage to wheat crop. For that, it is essential to find the pattern of how these diseases are transmitted across wheat cultivation and inspect the major factors that are influencing the formation of these patterns (Table 1).

Generally, government authorities keep record of report that farmers report on onset of the diseases and this information can be used as an input data for the tool proposed here. Another input approach can be crowd sourcing via a user friendly web-based or a mobile application. These data inputs can be sent to the tool and then eventually to a data storage using HTTP or REST API. According to Lau et al. [10], while installing weather monitoring sensors, they account for some errors or mistakes. Hence, we avoided the use of sensors as an input to the system as it will affect the result. Instead we obtain free satellite data available for public at minimal or no expense and also available at near-real time, example data from Sentinel satellites.

Vegetation maps can be generated using Sentinel-2B images and indices such as NDVI, EVI, NDWI and NDMI [7] can be performed to map the spatial distribution of vegetation and moisture present over a field. NDWI and NDMI map can help predict the spread depending on the water content near the disease reported area and the disease spreading rate depends on the density of vegetation which can be obtained from NDVI and EVI. Also, weather data e.g. precipitation, relative humidity at time of max temperature, (mean, min and max) temperatures, can be obtained from Sentinel-3 images or from local

weather stations. Another factor that plays an important role in spread is the wind. With the help of the intensity and direction parameters of the wind we can forecast the extent and duration up to which the disease can spread. The final output, the area of dispersed infection, can be produced in form of a interactive map.

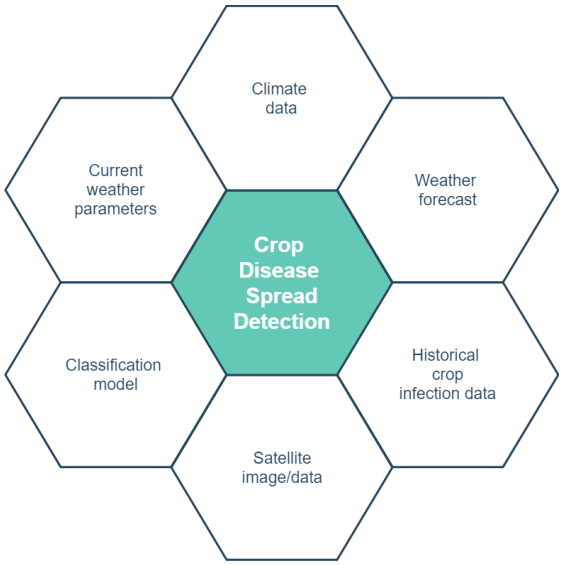


Figure 1: Major inputs of the CDSD tool.

Related work

Crop disease spread and impact

Wheat disease can become a serious problem for its cultivators and production rate if left untreated. Part of emerging problems related to plant disease is to figure out how these infections spread from plant to plant. Diseases that attack the leaves of a plant are primarily spread by wind, but they can also move to nearby plants by means of water from rain or irrigation. Transmission through the soil as a medium is also a possibility [17]. Hence it is important to identify reason of spread and try to prevent it at its origin.

Satellite Images and Machine learning

Satellites have been used for many years to measure crop damage due to natural calamities like hail-storm, flood, forest-fires, etc. and also to capture the damage caused by infections [11]. Studies are carried out to understand the environmental factors influencing the plant diseases using Geographic information science and satellite surveillance for diseases detection, spread and management [18]. Private-sector firms offer site-specific weather estimates, as well as disease and pest-warning advisory [1]. Our idea differs with such services, as we use input data to identify the evolving pattern of the spread of diseases. With the use of service like ERA5 by ECMWF, we can access many of the parameters like temperature, wind, rainfall, etc. that we assume to be related to disease spread [3]. Aeolus satellite provides wind data that is near real-time and hence can play an important role in the accuracy of the weather and climate predictions [1]. Thus, by these data we can identify the origin of the disease and spread of diseases in a particular direction.

Machine learning approaches like SVM, MLP and Random Forest are already used in recognising healthy and unhealthy plants [20]. Further colour based features like pixels, statistical features and Histogram of Oriented Gradients (HOG) have been used to classify healthy and unhealthy plant leaves. This is more enhanced when compared to the traditional methods for identifying plant diseases. Arivazhagan et al. [1], in their research, the shape and color features extracted from the affected region are given as input to the SVM classifier. Support Vector machine (SVM), a supervised machine learning algorithm, is being used in this paper for feature extraction and detection and to improve the accuracy of detection. Machine learning based detection, more specifically automatic recognition of plant diseases and exploring preventive methods for better recognition of plant diseases are carried out [14].

Ideally with the reach of satellite and local weather data this proposed tool can be operated in any region. For initial stages we propose to centre down our area of interest to North Rhine Westphalia (NRW), Germany. The wheat pathogen/disease risk tends to increase from 2001-2050 in NRW region [21,8]. For validating our concept, data of this federal state can be used for supervised learning, as the availability of data can be regular and in sufficient quantity.

Tool design/architecture

The proposed tool will have several different inputs that will come from different sources and will enter system at different levels. As you can see in the architecture diagram below, there are four different input points. The working flow can be roughly divided in three sections; first will be the reporting and validation phase. Fetching of archived and forecast data will be the next phase and it also involves passing this data to the main engine of tool. The final stage is to perform the prediction and display the result (Figure 1).

Table 1: List of parameter/variables used to predict the spread of diseases.
* Current and forecasted values.

Wind Speed *	Wind Direction *	Precipitation *
Air Temperature *	Soil Temperature *	Humidity *
Solar Radiation *	Pressure *	Land Use/Cover
Surface Runoff	Soil Type	

The input points for infection reporting and corresponding weather data are linked with the system storage component. There are two distinct sources for reporting inputs, either through a mobile application for farmers or web based user-interface. The archival data input point which is connected with storage component serves as an input to the operational component. This entry point will have three separate sources viz. well-maintained records of past disease spreads by the agricultural departments at regional, national or EU-level; second, similar historical records documented by different private institutes and organizations like agricultural research bodies or fertilizers/pesticides companies and finally, unstructured records like newspapers, case-study papers, journals, etc. For the first two sources the data needs to be migrated into a uniform format before entering into the archival component. While for unstructured data, we may need to create an automated script or a crawler to identify reports on crop infection or agricultural damages to put all such data under a structure. The operational component is responsible for the spread prediction. Input component responsible for fetching forecast data also sends its data to the operations sub-system.

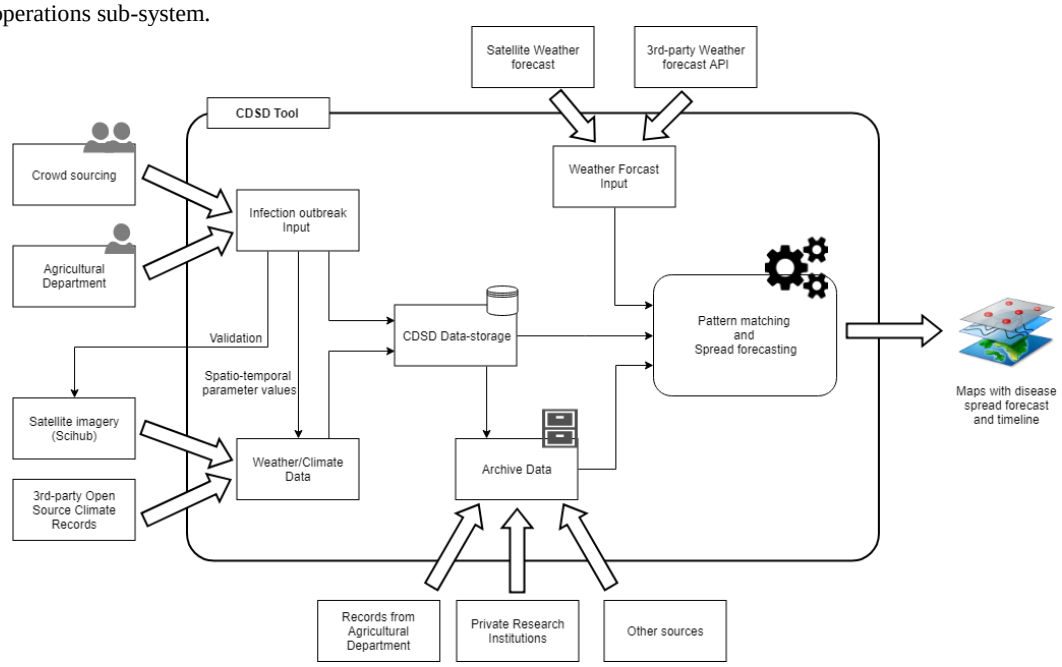


Figure 2: System architecture.

Processing

The overall processing is complex, distributed at multiple levels and interacting with different external interfaces. Also, there is a synchronous internal interaction happening between the different components of the system. Sub-processes, like using archiving samples and training the model can be done periodically in offline/background mode (Figure 2).

Validation

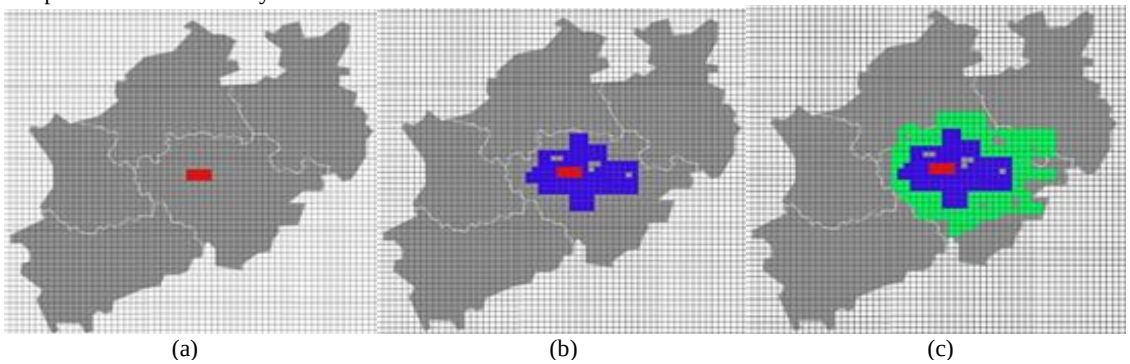
We propose to make an initial level validation before the main process starts. Here, we perform the validation of inputs related to occurrence of diseases reported through the input user interface. As input, the tool gets location and time period of disease occurrence, the satellite images corresponding to that geographical location and time interval are downloaded from the satellite data hub. Further, image processing operation developed for the identification of crop disease are employed here [6,12]. This will remove any false positives and re-confirm the existence of the infection on the wheat crop. One additional important part of the validation phase is to supply the model with the training data for the future which will make the results more accurate as the current data will be accurate and correlated strongly. The input data along with corresponding processed image can be stored in archive for lookup in the future. This can be used to maintain a log for identifying unreported infections using this tool in future.

Trigger
When using the tool in real time, the main prediction process can be triggered once there are enough number of infection incidents reported. This can be monitored by having threshold of ratio, 'x' reports per 'y' area (e.g. 20 incidents per 50 km²). Once the threshold is surpassed, a bounding box will be created based on the geographical coordinates of the reports. The centre of this bounding box will be the anchor point for determining the regions for checking the spread. After this, the model will operate in an area in the form of concentric circles with centre as the one determined for the bounding box. The radii for these circles can be incremented in fixed values (e.g. circles with radii as 10 km, 20 km, 30 km, etc.). The classification algorithm (mentioned in the next section) will be operated on each unit of area. This unit can be a single pixel or a group of pixels (square of side of 10 pixels) from the satellite image or an absolute area (e.g. 1 km²). Each circular region can be considered corresponding to a prediction for fixed number of days in future. For instance, 10 km circle corresponds to 1 week, 20 km circle with 2 weeks and so on.

Prediction
Though we proposed a wide array of parameters through past approaches and literature in this field, we need to validate the most necessary parameters that impact the spread of infection. We can also collaborate with experts in this field for their opinions and do further segregation of the important independent variables. Existing historic data can be used for carrying out a statistical analysis for each variable. The parameters with a P-value less that than 0.05 can be assumed as significant and the remaining can be discarded. Also the parameters can be weighted based on their significance to give more importance to significant variables than the lesser ones.

For predicting the spread, we propose to use classification algorithms for making the ML model. Among the many algorithms available, we suggest to try the Logistic Regression, Decision Tree and Random Forest as they suit best to do a classification on continuous and categorical data. Using the Receiver operating characteristics (ROC) plot the algorithm that gives most accurate output and least errors for a sample data set can be chosen as the best-fit main algorithm [9,4] and thus eliminating the rest [23]. Also, the Nearest Neighbour algorithm can be used to overcome the problem to fill the odd non-classified areas in a classified region that may appear during the evaluation of possible spread area (Figure 3b and Figure 3c, grey spots within the blue and green areas).

Once the parameters influencing the prediction are confirmed, we will derive the model equation based on significance of parameters and training dataset available. The equation will actually give a numerical value but that does not make it a regression model as based on this value we will categorized if the given unit area will be infected or not. For this, a threshold value will be referred to for comparison. This threshold will be calculated using the historical data where we are already aware of the values for input variables as well as the existence of the disease at a location. The above operation will be performed for every unit area in circular area surrounding the bounding box for given number of days. Once the first circle is completed we will get some areas that are true for spread while some will be false. Such areas will be identified as one continuous region and the false areas surrounded in the true region will be rounded as true using the hole filling algorithms from the image processing domain. The whole process will be repeated for all the unit areas in bigger circle with next step radius for spread detection. After checking the region for the biggest circle and greatest number of days in future, we get a graphical spatial timeline. This series of maps can be viewed by connecting the tool to any GIS vendor tool. This can be viewed by the researchers and employees at agricultural department and can be published through a GeoServer as a service so that the spread can be viewed by the end users like farmers.



(a) (b) (c)

Figure 3: (a) Red area indicates bounding box of locations of infection reported. (b) Blue area highlighting the prediction of area under infection spread after 1 week. (c) Green area highlighting the prediction of area under infection spread after 2 weeks.

We need to be sure of certain points while operating this prediction tool. The output can be reasonable only if there is certainty of correctness to the inputs and their sources. If we cannot guarantee the correctness of the data then it should be discarded. Out of the available data, generally 70% is used for training the model while the remaining 30% is used for testing. In case the data prior to execution is less, then the proportion can be shifted to 80%-20%.

Limitations and Challenges

The overall scope of this concept is extremely big and it involves multiple levels of operations. There are challenges that can surface during the actual development and implementation of this system. As we did not get a chance to build this idea in reality, below we have tried to mention few of the difficulties that may occur and needs to be taken care of.

- False reporting of data by farmers due to incorrect guess of a disease and even lack of understanding of how to handle and operate the application is very much possible.
- Quality of data can always be a challenge when it comes to use of forecasted data or data obtained from crowd-sourcing.
- Though the new satellite services promise to provide almost real time images, there are possibilities of delay due to technical failures or malfunctioning of any satellite components or ground stations.
- Availability of historical records for occurrence of the specific disease for a particular region is still a major doubt.
- Though the prediction is based on validated input that, there can be a degree of error associated which can be analyzed and cancelled out only after the real testing in field.
- The final prediction of disease spreading is based on multiple factor, weather entities like temperature, wind, precipitation being few of them. These input attributes are themselves forecasted and may get altered. Hence the final outcome is much more prone to uncertainty.

Future work

The development and implementation of the proposed methods will be the first step going forward. As mentioned in the limitations, if sufficient data from the past is not available then work must be done to gather data in a structured format. In future other possible input parameters that may influence the spread pattern can be included in the system. The tool proposed here can be used for other crops and diseases other than wheat and stem rust respectively. Researchers have developed a plant disease recognition model using deep convolutional neural networks (CNN) [19]. The CNN database can be used in the existing system for differentiating various diseases and help the farmers be more aware of the diseases generated in their crops.

Conclusion

The idea when achieved can hold great ability to minimize the loss of food crop. The preventive measure presented here can facilitate lesser loss to crop yield by giving time to the authorities, organizations and the farmers to act on the future losses. The idea is still in maturing days but our group have made sure to make a thorough research behind every aspect that we have presented here. All the problems, possibilities and concepts are studied in detail as per our technical and scientific capabilities.

References

1. Aeolus (no date). ESA's wind mission. https://directory.eoportal.org/web/eoportal/satellite-missions/a/aeolus 12
2. Arivazhagan, S., Shebiah, R. N., Ananthi, S., & Varthini, S. V. (2013). Detection of unhealthy region of plant leaves and classification of plant leaf diseases using texture features. Agricultural Engineering International: CIGR Journal, 15(1), 211-217.
3. Climate Data Store, https://cds.climate.copernicus.eu/cdsapp#!/dataset/era5-hourly-data-on-single-levels-from-2000-to-2017?tab=overview
4. Decision tree or Logistic regression. In Stack Exchange, Retrievd 16:20, September 11,2018 from https://datascience.stackexchange.com/questions/6048/decision-tree-or-logistic-regression
5. Gleason, M. L., Duttweiler, K. B., Batzer, J. C., Taylor, S. E., Sentelhas, P. C., Monteiro, J. E. B. A., & Gillespie, T. J. (2008). Obtaining weather data for input to crop disease-warning systems: leaf wetness duration as a case study. Scientia Agricola, 65(SPE), 76-87.
6. Huang, W., Luo, J., Zhang, J., Zhao, J., Zhao, C., Wang, J., ... & Du, S. (2012). Crop disease and pest monitoring by remote sensing. In Remote Sensing-Applications. InTech.
7. Indices, https://www.sentinel-hub.com/develop/documentation/eo_products/Sentinel2EOproducts 11
8. Juroszek, P., & von Tiedemann, A. (2013). Climate change and potential future risks through wheat diseases: a review. European Journal of Plant Pathology, 136(1), 21-33.
9. Lan, H. (2018). Decision trees and random forest for regression pt1 and classification. Published in Medium, Towards Data, Retrieved 11:03, September 10, 2018 https://towardsdatascience.com/decision-trees-and-random-forests-for-classification-and-regression-pt-1-dbb65a458df
10. Lau, Y. F., Gleason, M. L., Zriba, N., Taylor, S. E., & Hinz, P. N. (2000). Effects of coating, deployment angle, and compass orientation on performance of electronic wetness sensors during dew periods. Plant Disease, 84(2), 192-197.
11. Lowe, A., Harrison, N., & French, A. P. (2017). Hyperspectral image analysis techniques for the detection and classification of the early onset of plant disease and stress. Plant methods, 13(1), 80.
12. Mahlein, A. K., Kuska, M. T., Thomas, S., Bohnenkamp, D., Alisaac, E., Behmann, J., ... & Kersting, K. (2017). Plant disease detection by hyperspectral imaging: from the lab to the field. Advances in Animal Biosciences, 8(2), 238-243.
13. McIntosh, R. A. (2009). History and status of the wheat rusts. In Proceedings of the 2009 Technical Workshop Borlaug Global Rust Initiative, Cd. Obregon, Sonora, Mexico, March (pp. 17-20).
14. Megha, S., Niveditha, C. R., SowmyaShree, N., & Vidhya, K. (2017). Image Processing System for Plant Disease Identification by Using FCM-Clustering Technique. International Journal of Advance Research, Ideas and Innovations in Technology, 3(2), 445-449.

15. News of wheat diseases. https://www.independent.co.uk/news/uk/home-news/wheat-rust-the-fungal-disease-that-threatens-to-destroy-the-world-crop-9271485.html; https://www.reuters.com/article/us-wheat-pest-europe/scientists-fear-resurgence-of-devastating-wheat-disease-in-britain-europe-idUSKBN1FS1C5
16. Savary, S., Ficke, A., Aubertot, J. N., & Hollier, C. (2012). Crop losses due to diseases and their implications for global food production losses and food security.
17. Science Day Friday. How Do Infections Spread In Plants? https://www.sciencefriday.com/educational-resources/how-do-diseases-spread-between-plants/ 15
18. Singh, R., Ranjan, K., & Verma, H. (2015). Satellite imaging and surveillance of infectious diseases. Journal of Tropical Diseases & Public Health.
19. Sladojevic, S., Arsenovic, M., Anderla, A., Culibrk, D., & Stefanovic, D. (2016). Deep neural networks based recognition of plant diseases by leaf image classification. Computational intelligence and neuroscience, 2016.
20. ur Rahman, H., Ch, N. J., Manzoor, S., Najeeb, F., Siddique, M. Y., & Khan, R. A. (2017). A comparative analysis of machine learning approaches for plant disease identification. Advancements in Life Sciences, 4(4), 120-126.
21. Volk, T., Epke, K., Gerstner, V., Leuthner, C., Rotterdam, A., Johnen, A., & Richthnfen, J. S. V. (2010). Klimawandel in Nordrhein-Westfalen–Auswirkungen auf Schädlinge und Pilzkrankheiten wichtiger Ackerbaukulturen. Münster: proPlant GmbH.
22. Wageningen University & Research. European MARS crop yield forecasting system, https://www.wur.nl/en/show/Agricultural-crop-monitoring.htm
23. Wikipedia contributors. (2018, August 29). Receiver operating characteristic. https://en.wikipedia.org/w/index.php?title=Receiver_operating_characteristic&oldid=857016426