*Cilantro*: An Artificial Intelligence and Graph Theory Based Novel Methodology to Model the Spread of Infectious Potato Diseases

## 

## Abstract

The risk of any pathogenic infection like *Phytophthora Infestans* in potatoes and the speed at which it can spread depend on several atmospheric, biological, genomic and geographic factors. In this paper, we are presenting *Cilantro*, an AI and graph theory based solution that we have developed to detect the infection of certain pathogens in potatoes at an early stage, model the spread of these diseases from different perspectives and analyze the spread pattern. The novel AI algorithm of *Cilantro* ensures high precision risk estimates of *Phytophthora Infestans* outbreak based on environmental and geographical parameters. The algorithm uses appropriate crowd sourced information regarding such infestations. *Cilantro*, unlike other applications in this domain, has the ability to dynamically integrate new findings and contextual disease recognition alerts from edge servers as well as mobile devices using artificial neural networks for disease detection and actual pathological and epidemiological data collected from multiple laboratories of agricultural extension services. We believe that *Cilantro* will help farmers take better precautions against disastrous diseases like late blight that often cause huge financial and social challenges in many countries, thereby, protecting their harvest with the help of early detection, prevention, and treatment.

**Keywords**

## Introduction

For years, late blight infection has been the subject of many empirical forecasting models [1-8] as well as mechanistic simulation models [2, 9-21]. Empirical forecasting systems are often derived from *Blitecast* [23], and are based on the accumulation of late blight risk units under daily temperature and humidity conditions summarized from hourly data [4-6,28,13]. However, these models often are unsuitable for commercial implementations involving large-scale data sets derived from real-time user reports. The *Blightcast* model also has no provision for transfer learning or ergonomics and was originally developed with a single disease in mind. Studies have shown that simple empirical models generally fail to properly address the hidden risk of oospores that often remain dormant in the soil for more than a year [22].

Implementing real-time user reports is an important feature to include, especially in models that predict the spread of agricultural diseases in developing nations, where accurate weather and remote-sensing data might not always be available. Furthermore, accounting for expected as well as random changes in pathogen characteristics and fungicide resistance due to climate change [22,25-27] is crucial in any modern epidemiological modeling system.

## Problem Statement

Late Blight in potatoes, caused by *P. Infestans* can quickly spread across geographically connected cultivation fields through infected tubers, making it incredibly difficult to monitor or control. While traditional machine learning methods are capable of recognizing the presence of late blight in farms, they fail to provide accurate early warnings and active monitoring guidelines to potato growers and supply chain managers. An average farmer cannot rely on these technologies since a possible blight outbreak needs expensive and labor intensive treatment procedures. On the other hand, a false negative alert while calculating the risk of blight in a potato tuber can severely decimate the yield capacity of a single cropping season, making the need for an accurate early warning and risk profiling methodology the need of the hour.

## Proposed Solution

## Overview

To tackle the above-mentioned problems, we are introducing a new AI and Graph-based solution for monitoring and analyzing Blight outbreaks. This analytical model consists of primarily three components: A graph-based risk analysis system, an AI and Computer Vision-based diagnostic system, and a comprehensive user reporting system for dynamically recalibrating the graph-based model.

### Description of the Graph-Based Model

We subdivide a given geographical region into some elementary zones. Each of these elementary zones gets treated as a node in a unidirectional, weighted graph. The entire landmass is divided into n equally sized land blocks. Each of these blocks gets represented by a single node in a graph with the following properties and features:

1. Each block is connected to eight other neighboring blocks by direct land connection.
2. One block can be connected to any other block in the following ways:
   1. Air Route
   2. Road
   3. Rail
   4. Sea Route
   5. Sea Current
   6. Air current/Air Stream
   7. Foot trail
3. Each block has a risk factor R associated with it. R is the probability of finding a case of the disease in that location.

A node L1 can be connected to any other node L2 by a directed edge, the weight of which depends on the Vector Score V(L1,L2). V can be represented as,

V(L1,L2) = P(L1|L2)

Where P(L1) is the probability of finding the disease in node/block L1. P(L1|L2) is the probability of finding the disease, provided the disease has been found in node/block L2. The formula for calculating the edge weight can be as follows.

V = Vn .D(Vn,t)

Vn = [ a w i h ]

Where,

a = P(L1|L2) given the mode of transmission is air at the exact moment the disease is first detected in L2

w = P(L1|L2) given the mode of transmission is water at the exact moment the disease is first detected in L2

i = P(L1|L2) given insects act as a vector at the exact moment the disease is first detected in L2

h = P(L1|L2) given humans act as a vector at the exact moment the disease is first detected in L2

Vn is also called the Transport Factor. It is an intrinsic, time-invariant property of a particular edge.

The matrix is also called Decay Vector D(Y,t) for the vector Y. It specifies how the risk of each transmission medium depends on time. Each fn(t) has the general form

fn(t) = + c

Where,

= constant that depends on the vector and disease

= constant that depends on the government policies and quarantine methods

= constant that depends on the disease

c = positive constant that depends on the disease

For V(L1,L2), Y = Vn. Each node also has a Susceptibility Score S(L1,m). Where S can be represented as,

S = (E . ) .

Sn = [ T r u v ]

Where,

T = Average temperature of the node/block L1 in Kelvin for the month m

r = Average rainfall recorded in the node/block L1 in mm for the month m

u = Average relative humidity in the node/block L1 in % for the month m

v = Average irradiance of solar radiation with wavelength 400nm - 1900nm received per meter square in the node/block L1 measured in W/m2 for the month m

E = A square matrix of size dim(Sn)xdim(Sn) that represents the dependency of the disease on a particular environmental factor.

From the above, it can be deduced that S(L1,m) can also be represented as,

S(L1,m) = P(L1|t)

Where,

P(L1|t) = Probability of finding the disease at L1 at time t.

The risk can be formulated as

The expression is called *Recovery Coefficient* and it specifies the rate at which the effect of the disease dies off (in other words diminishing gradient)

= number of nodes connected to the node Lk by a directed edge to Lk.

The constants depend on the particular disease.

= the number of blocks or nodes.

= Memory Coefficient which represents the fact that the risk of a disease reappearing is always finite and non-zero.

Since the graph is not fully connected, the adjacency matrix corresponding to this graph is a sparse matrix. Thus, traditional sparse matrix representations are adequate in representing the graph. The Risk Factor of a given node only states the probability of finding the disease given the Risk Factor of nodes/blocks connected to it by a directed, weighted edge from the other nodes.

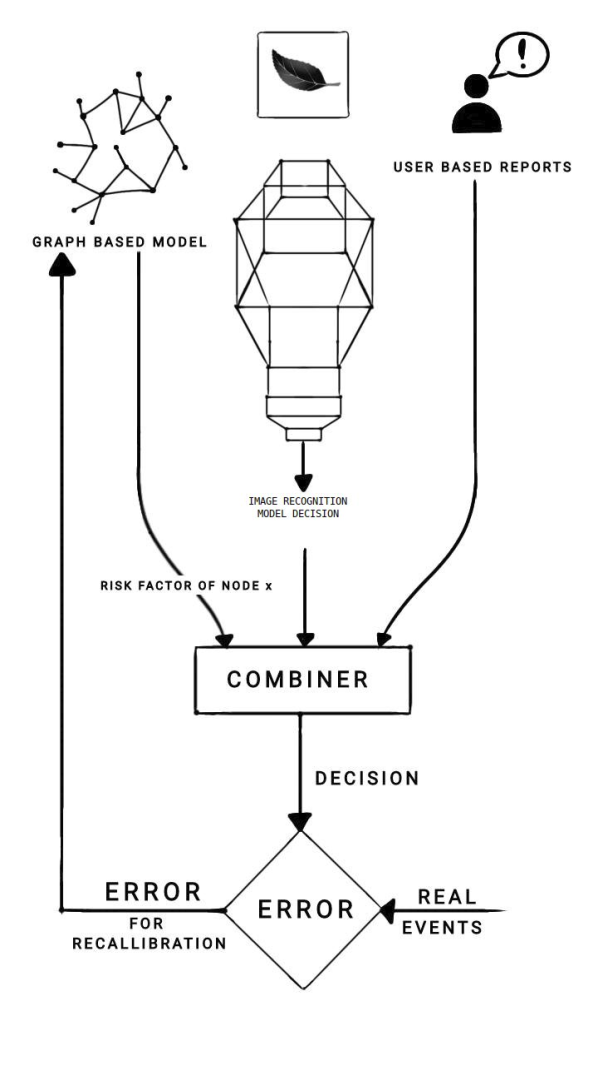
### Algorithm to model the spread of diseases

The algorithm to simulate disease spread is:

1. The nodes, in which the disease has already been found, the Risk Factor is initialized to a random value between 0.9 and 1.
2. Initialize the adjacency matrix.
3. Initialize time as t=t0.
4. If t<tf, continue. Else, Jump to end
5. t := t + dt
6. Recompute the adjacency matrix
7. Compute the average Risk Factor among the nodes with a non-zero Risk Factor.
8. For all the nodes having higher than average risk factors, increase the Risk Factor by multiplying the Risk Factor with (1 + j), where j is a positive constant between 0 and 1.
9. End.

The average Risk Factor is treated as the Universal Risk Factor Ur

A collection of such graphs can be mapped into a single graph by treating each individual graph as a node and its corresponding Universal Risk Factor as its Risk Factor. In this case, the whole system can be treated as a fractal. Scaling up this system would also be highly efficient with this method.

*Figure 1: Description and flowchart of Cilantro architecture*

### User-Based Reports

Reports collected from the user base are extremely important in this model. *Cilantro* will collect reports from existing users of the technology stack and use the gathered information both as a starting point for the graph-based model and generate errors to calibrate the model every 24 hours. The user reports are mapped into a 10-point severity scale and are represented as a sparse matrix and normalized.

## Experimental Results and Analysis

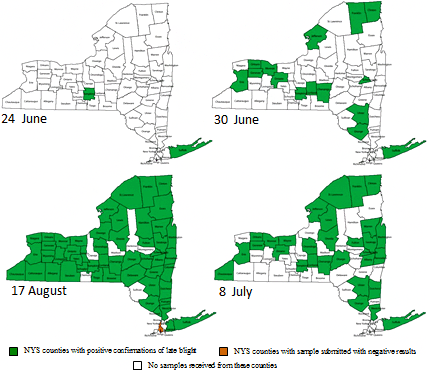
## Experiments

To gauge the validity of the model, we consider the spread of the pandemic between the dates June 24, 2009, and July 8, 2009, [23] for 59 different counties in the state of New York, US. We tabulate the spread of the disease into a data frame with counties as rows and dates as columns. The cells have two values: 1 if samples from the county tested positive for blight; 0 if the samples tested negative for blight. The value can be fractional if no samples are available for that date. For each county, we compute the edge weight by calculating the geographical distance between each pair of counties. The resultant 59 x 59 distance matrix is fed into the model. The probability output from the model is compared with the real data obtained. If the probability is reasonably high, beyond a certain threshold, the county is declared infected and the resulting confusion matrix is compared.

## Dataset Used

The data used for the model falls into three categories:

1. Epidemiological data: Constants describing how the oospores and zoospores of *P. Infestans* spread.
2. Weather Information: Weather data, in this case, refers to the wind vector, surface temperature, Relative Humidity, daily minimum and maximum temperature in the state of New York for the summer of 2009 from the JRA-55 dataset. The information is represented as a vector map that represents hourly wind, precipitation, temperature, and wind direction information in a JSON file, which is then decoded using a custom, proprietary Python script.
3. User reports: User reports are generated randomly following a gaussian distribution and a decay function and in another case, following a uniform distribution. The user reports in a real scenario can be perfectly uniform or heavily biased. However, in an experimental stage, since we are testing the model’s ability to analyze the outbreak and give timely forecasts, a simulated user reporting system will suffice.
4. Real-world disease spread data: As described in section 4.1, we have obtained disease spread information from the National Plant Diagnostic Network (NPDN) and have tabulated the data using tesseract and positional OCR.

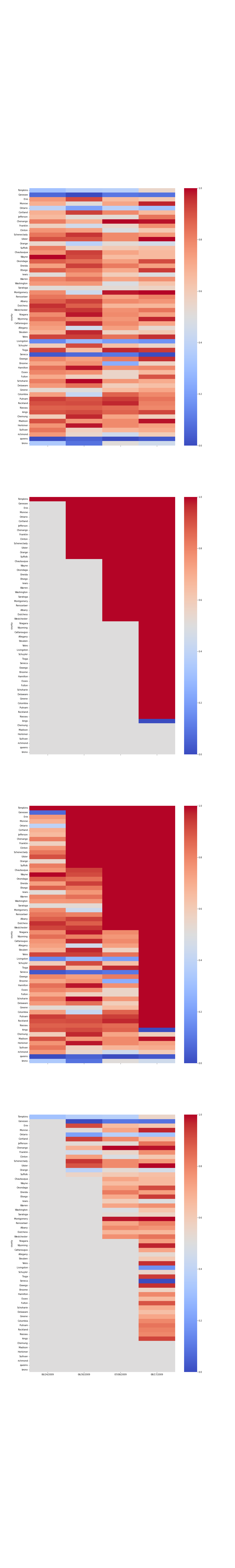


*Figure 3: Counties in New York which had reported late blight (in a garden center or in the field) during the summer of 2009. [23]*

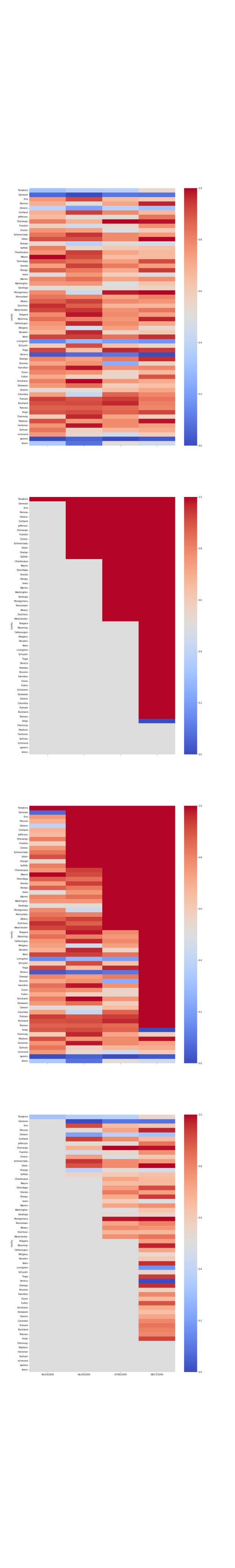
The dataset is converted into a graph-assisted spatio temporal representation by first obtaining the latitudes and longitudes for the various counties represented on the map. Then the adjacency matrix is computed using the geographical distance between each county and the relative density of road interchanges within the county. Since the oospores and zoospores can spread through infected farm equipment, this is an important factor that must be incorporated into the dataset, to accurately predict the spread of blight. The temporal part of the dataset is generated automatically using a custom Python script and the resultant dataset is fed into the model.

## Results and Discussion

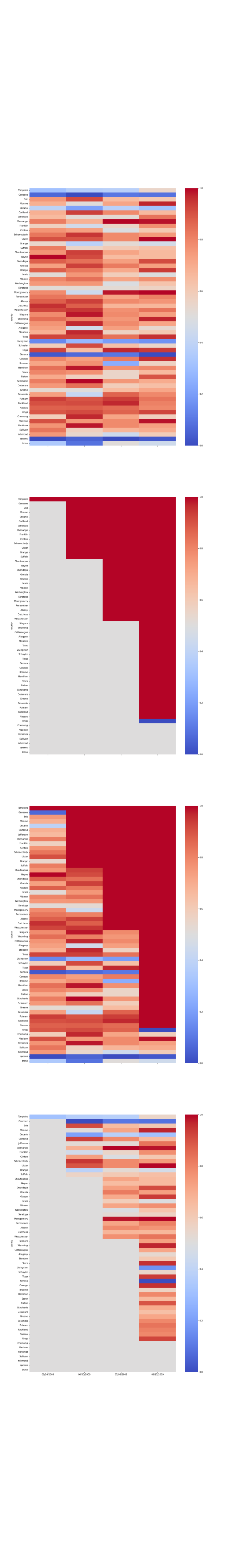
Simulating the disease spread, we clearly see an increase in risk value just before a disease is detected. The risk value decreases after disease detection, in line with the hypothesis presented in this paper.

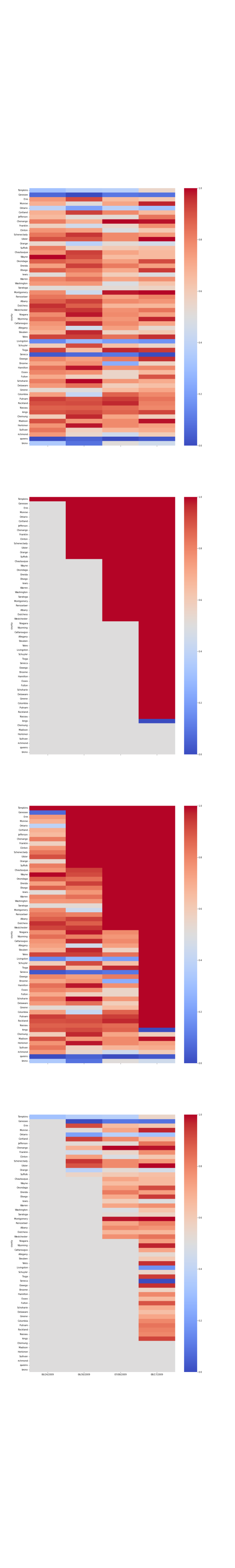


*Figure 1: The above diagram represent the heat map of the ground truth verses the predicted risk by the cilantro engine. Lighter colours shows the higher risk of disease and dark colour shows the lower risk of disease.*



*Figure 2: In the diagram the risk is 0.5 when there is no report available, and the risk is 0 when the report is false which means no blight is detected and it is 1 when blight is detected.  
The counties are grouped together for easier visualisation and analysis and as we can see there are several county's which got a positive report or where blight is detected together which means blight spread to many counties at the same time.  
At the topmost graph it represents the risk which is how much is the probability of blight appearing in each county.*



*Figure 3: This diagram is the combination of both the above graphs and that is the actual output that the engine will give. only where the ground truth is 0.5 which is where the report is not available the cilantro engine will step in and provide “risk” value.  
Here “risk” is the probability of blight occurring in that position at a certain date as described in the above formulas.*

*Figure 4: Here we look at the risk value that the machine predicts after the blight has been detected.  
We observe how the machine treats the events after days after which blight is detected and here, we saw lower risk which is quite intuitive since after blight is detected the risk of blight being detected again is lowered because of the aggressive treatment that the farmer provides.  
For the ground truth we have assumed once the blight is detected the risk is highest among 0 and 1 which is 1. If blight is definitely not detected which is a negative report from our lab then we have decided the risk value to be 0 and if there is no report available or we do not have any information about that area we label the risk value as a superimposition of the two different risk values namely the risk value becomes an average of the two which is 0 plus 1 divided by 2 which gives us 0.5. In an actual scenario only when the ground truth is 0.5 will cilantro engine will step in and provide a risk value.  
The cilantro engine will work or will cease operation once a lab provides a definitive report.*

1. Analysis of Results

Now we can analyse these results to demonstrate the robustness of *Cilantro*. In this section, we will pick up a few counties to show the precision of our risk prediction. In county Tompkins, Geneseethe engine predicted that there would be no blight outbreak for the next few weeks. The actual field report indeed supports that. For Erie, Schenectady the risk value increases as the blight gets detected and also gradually decreases along the week after the detection because of the aggressive treatment by the farmers. From Cortland to Ulster County the risk initially dipped in the first week and then increased and only then there was a detectable case blight outbreak according to the report. The predicted risk value went down marginally after the initial detection.

From Suffolk to Yates and Oswego to Kings, we can see that initially no blight infection was detected. *Cilantro*’s ML engine predicted there would be an increase in the risk value. The field data supported this prediction. In counties such as Schuyler, Tioga and Monroe, initially no blight infection was detected. However, at a later stage farmers observed an outbreak. *Cilantro* correctly predicted this risk based on the available parameters.

The correlation between high-risk value and positive cases in respective counties indicate the correctness of the model. To simplify the calculation, the decay function of f(x) = 0.6667x is chosen as the decay value, while the vectors and atmospheric conditions are treated as independent random variables.

## Future Work

In future, we plan to extend this research work by concentrating ona few different techniques for fine-tuning this model by further quantifying the severity scale of various plant diseases. This quantification process is expected to increase the accuracy of the model and decrease its reliance on user reports. This will make the model suitable for precision treatment and will enable error detection as well as recalibration using remote sensing satellites. Down the line, we want to focus on creating novel methods for disease control, prevention and avoidance.

## Conclusion

In this paper, we have presented *Cilantro*, a novel graph-based, scalable artificial intelligence algorithm that can predict the risk of Blight outbreaks in a particular location. *Cilantro* is scalable and possesses a significant advantage over other commercial software in the domain of agricultural pathology, due to its ability to take into account user reports and atmospheric data while modelling the spread of Blight. We have tested *Cilantro* by using it to predict the 2009 Blight outbreak in New York, USA. *Cilantro* correctly predicted the outbreak by indicating an increase in risk values in counties where positive Blight cases were found. Further testing using higher resolution, comprehensive spatio-temporal datasets will further prove the accuracy and ability of this model.

## References

1. Baker, K.; Lake, T.; Benston, S.; Trenary, R.; Wharton, P.; Duynslager, L.; Kirk, W. Improved weather-based late blight risk management: Comparing models with a ten year forecast archive. J. Agric. Sci. 2015, 153, 245–256. [CrossRef]
2. Fry, W.; Apple, A.; Bruhn, J. Evaluation of potato late blight forecasts modified to incorporate host resistance and fungicide weathering. Phytopathology 1983, 73, 1054–1059. [CrossRef]
3. Garcia, B.I.L.; Sentelhas, P.C.; Tapia, L.R.; Sparovek, G. Climatic risk for potato late blight in the Andes region of Venezuela. Sci. Agric. 2008, 65, 32–39. [CrossRef]
4. Grünwald, N.J.; Montes, G.R.; Saldaña, H.L.; Covarrubias, O.R.; Fry, W.E. Potato late blight management in the Toluca Valley: Field validation of SimCast modified for cultivars with high field resistance. Plant Dis. 2002, 86, 1163–1168. [CrossRef] [PubMed]
5. Hijmans, R.J.; Forbes, G.; Walker, T. Estimating the global severity of potato late blight with GIS-linked disease forecast models. Plant Pathol. 2000, 49, 697–705. [CrossRef]
6. Iglesias, I.; Escuredo, O.; Seijo, C.; Méndez, J. Phytophthora infestans prediction for a potato crop. Am. J. Potato Res. 2010, 87, 32–40. [CrossRef]
7. Johnson, D.A.; Cummings, T.F.; Fox, A.D. Accuracy of rain forecasts for use in scheduling late blight management tactics in the Columbia Basin of Washington and Oregon. Plant Dis. 2015, 99, 683–690. [CrossRef] Pathogens 2020, 9, 659 18 of 20
8. Kaukoranta, T. Impact of global warming on potato late blight: Risk, yield loss and control. Agric. Food Sci. 1996, 5, 311–327. [CrossRef]
9. Apel, H.; Paudyal, M.; Richter, O. Evaluation of treatment strategies of the late blight Phytophthora infestans in Nepal by population dynamics modeling. Environ. Model. Softw. 2003, 18, 355–364. [CrossRef]
10. Aylor, D.E.; Fry, W.E.; Mayton, H.; Andrade-Piedra, J.L. Quantifying the rate of release and escape of Phytophthora infestans sporangia from a potato canopy. Phytopathology 2001, 91, 1189–1196. [CrossRef] [PubMed]
11. Bruhn, J.; Fry, W. Analysis of potato late blight epidemiology by simulation modeling. Phytopathology 1981, 71, 612–616. [CrossRef]
12. Fall, M.; Van der Heyden, H.; Brodeur, L.; Leclerc, Y.; Moreau, G.; Carisse, O. Spatiotemporal variation in airborne sporangia of Phytophthora infestans: Characterization and initiatives towards improving potato late blight risk estimation. Plant Pathol. 2015, 64, 178–190. [CrossRef]
13. Henshall, W.; Shtienberg, D.; Beresford, R. A new potato late blight disease prediction model and its comparison with two previous models. N. Z. Plant Prot. 2006, 59, 150–154. [CrossRef]
14. Michaelides, S. A simulation model of the fungus Phytophthora infestans (Mont) De Bary. Ecol. Model. 1985, 28, 121–137. [CrossRef]
15. Raymundo, R.; Andrade-Piedra, J.; Juárez, H.; Forbes, G.; Hijmans, R.J. Towards an integrated and universal cropdisease model for potato late blight. In Late Blight: Managing the Global Threat, Proceedings of Global Initiative on Late Blight (GILB) Conference, Hamburg, Germany, 11–13 July 2002; Lizárraga, C., Ed.; International Potato Center (CIP): Lima, Peru, 2002; pp. 77–82.
16. Johnson, A.C.S.; Frost, K.E.; Rouse, D.I.; Gevens, A.J. Effect of temperature on growth and sporulation of US-22, US-23, and US-24 clonal lineages of Phytophthora infestans and implications for late blight epidemiology. Phytopathology 2015, 105, 449–459. [CrossRef]
17. Shtienberg, D.; Doster, M.; Pelletier, J.; Fry, W. Use of simulation models to develop a low-risk strategy to suppress early and late blight in potato foliage. Phytopathology 1989, 79, 590–595. [CrossRef]
18. Skelsey, P.; Kessel, G.; Holtslag, A.; Moene, A.; Van Der Werf, W. Regional spore dispersal as a factor in disease risk warnings for potato late blight: A proof of concept. Agric. For. Meteorol. 2009, 149, 419–430. [CrossRef]
19. Skelsey, P.; Kessel, G.; Rossing, W.; Van Der Werf, W. Parameterization and evaluation of a spatiotemporal model of the potato late blight pathosystem. Phytopathology 2009, 99, 290–300. [CrossRef]
20. Small, I.M.; Joseph, L.; Fry, W.E. Evaluation of the Blight Pro decision support system for management of potato late blight using computer simulation and field validation. Phytopathology 2015, 105, 1545–1554. [CrossRef]
21. Van Oijen, M. Selection and use of a mathematical model to evaluate components of resistance to Phytophthora infestans in potato. Neth. J. Plant Pathol. 1992, 98, 192–202. [CrossRef]
22. Narouei-Khandan, Hossein A., et al. "BLIGHTSIM: A new potato late blight model simulating the response of Phytophthora infestans to diurnal temperature and humidity fluctuations in relation to climate change." *Pathogens* 9.8 (2020): 659.
23. Fry, W. E., et al. "The 2009 Late blight pandemic in Eastern USA." (2012).
24. 24F. Wang, Z. Xuan, Z. Zhen, K. Li, T. Wang, and M. Shi, “A day-ahead PV power forecasting method based on LSTM-RNN model and time correlation modification under partial daily pattern prediction framework,” *Energy Convers. Manag.*, vol. 212, no. 112766, p. 112766, 2020. <https://doi.org/10.1016/j.enconman.2020.112766>
25. 25A. R. Blaustein, S. C. Walls, B. A. Bancroft, J. J. Lawler, C. L. Searle, and S. S. Gervasi, “Direct and indirect effects of climate change on amphibian populations,” *Diversity (Basel)*, vol. 2, no. 2, pp. 281–313, 2010. [**https://doi.org/10.3390/d2020281**](https://doi.org/10.3390/d2020281)
26. 26I. Kurane, “The effect of global warming on infectious diseases,” *Osong Public Health Res. Perspect.*, vol. 1, no. 1, pp. 4–9, 2010. DOI : <https://dx.doi.org/10.1016%2Fj.phrp.2010.12.004>
27. 27M. Matzrafi, “Climate change exacerbates pest damage through reduced pesticide efficacy,” *Pest Manag. Sci.*, vol. 75, no. 1, pp. 9–13, 2019. DOI: <https://doi.org/10.1002/ps.5121>
28. 28“Blight,” *Com.au*. [Online]. Available: https://www.yates.com.au/plants/problem-solver/diseases/blight/. [Accessed: 26-Sep-2021].
29. 29N. Alam, “Late Blight of potato *Phytophthora infestans*. - name: azmeagri study,” *Nazmeagristudy.com*, 10-Jun-2021. [Online]. Available: www.nazmeagristudy.com/late-blight-of-potato-phytopthora-infestans/. [Accessed: 26-Sep-2021].
30. 30K. J. Mutuku, M. J. Joseph, N. M. Wabomba, and K. A. Mwangi, “Efficacy of *Trichoderma asperellum* seed treatment and Ridomil® application in managing late blight on potato,” *World J. Agric. Res.*, vol. 9, no. 2, pp. 42–52, 2021. DOI: 10.12691/wjar-9-2-1
31. 31“Late blight of tomato – NOVIC daughter trial: Tomatoes,” *Evergreen.edu*. [Online]. Available: http://blogs.evergreen.edu/fieldstudy-patrick/late-blight-of-tomato-a-brief-history-of-the-disease-and-its-impact-and-current-knowledge-of-resistance-with-a-focus-in-western-washington/. [Accessed: 01-Oct-2021]
32. 32J. B. Ristaino, “Tracking historic migrations of the Irish potato famine pathogen, *Phytophthorainfestans*,” *Microbes Infect.*, vol. 4, no. 13, pp. 1369–1377, 2002. DOI: [10.1016/S1286-4579(02)00010-2](http://dx.doi.org/10.1016/S1286-4579(02)00010-2)