*Cilantro*: An Artificial Intelligence Powered Novel Graph-Based Data Analytics Engine to Model Infection Spread and Predict Future Outbreaks of Critical Potato Diseases

Abstract

The occurrence and spread of any pathogenic infection, such as *Phytophthora Infestans* in potatoes, primarily depend on numerous climatic, biological, genetic, and geographic parameters. Over the years, several reports collected from various parts of the world have demonstrated how the infection patterns and their rapid invasions can vary widely. In this paper, we are introducing *Cilantro*, a novel graph theory-based artificial intelligence solution we designed and developed to model the spread of such pathogen infections in potatoes, examine the pattern of dissemination and predict the future outbreaks of such infections under different circumstances and conditions. Using environmental and geographic data as well as pertinent crowd-sourced information about such infections, *Cilantro*’s ML-engine ensures high accuracy risk estimation of future *Phytophthora Infestans* outbreaks. As a robust and scalable solution, *Cilantro* continuously updates its risk ratings (prediction indices) based on various real life test reports collected from fields and employs a density directed, weighted graph representation of a blight-prone area. Unlike other applications in this field, *Cilantro* uses artificial neural networks for disease detection and dynamically integrates pathological results collected from multiple laboratories of agricultural extension services with contextual disease recognition alerts from edge servers, mobile devices, and desktop computers. We believe that *Cilantro* will assist farmers in better protecting their harvest through early detection, prevention, and treatment of devastating diseases like late blight, that frequently cause significant financial and social challenges in many countries.

**Keywords**

1. Introduction

For many years, both mechanistic simulation models [2, 9-21] and empirical forecasting models [1-8] have been used to study late blight infection. The accumulation of late blight risk units under daily temperature and humidity conditions summed up from hourly data [4-6, 28, 13] is the foundation of empirical forecasting systems, which are frequently derived from *Blitecast* [23]. When hourly meteorological data and artificial neural network models using LSTM-RNN [24] are used together to generate these predictions, the forecasts become more comprehensive and precise [1].

These models, however, often are not suitable for commercial applications incorporating big data sets collected from multiple sources including real-time field data and user feedback. The *Blightcast* concept was initially created with a specific illness in mind and did not include advanced analytical techniques such as transfer learning or ergonomics. Through multiple experiments and observations, the research community has understood that simple empirical models often fail to adequately address the hidden danger of oospores, which frequently remain latent in the soil for more than a year, according to studies [22].

It is crucial to provide real-time user reports, particularly in models that predict the development of agricultural diseases in developing countries where precise meteorological and remote-sensing data might not always be accessible. In addition, due to the increase in the frequency and severity of diseases due to factors such as global warming [27] and mutations of pathogens, accounting for expected as well as random unprecedented changes in pathogen characteristics and fungicide resistance due to climate change [22,25-27] have become essential in any modern epidemiological modelling system.

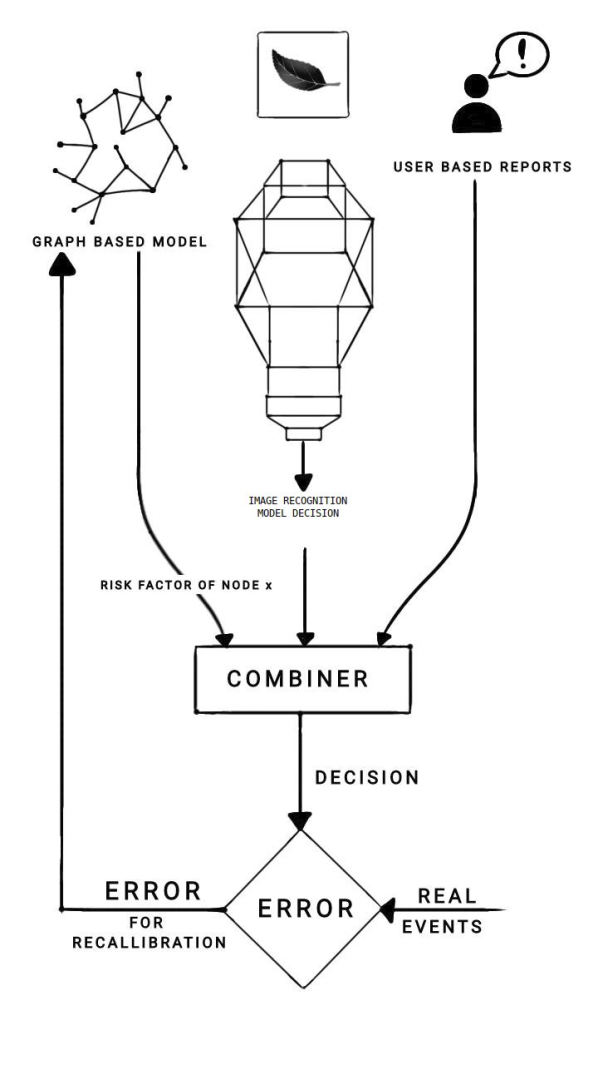
In this paper, we offer an innovative, graph-based, adaptive artificial intelligence powered solution that detects occurrences of some specific diseases, analyses the spread patterns of those ailments and predicts future outbreaks of such infections using multiple data sources including real-time user reports of plant infection outbreaks. This engine is also capable of dynamically modifying and calibrating itself to produce precise and contextual forecasts of impending blight epidemics acorss a large area.

1. Problem Statement

It is extremely challenging to monitor and control *Late Blight* infections in potatoes caused by *P*. *Infestans* using manual techniques of epidemiological monitoring because diseased tubers can disseminate the agents swiftly over geographically contiguous agricultural areas. While traditional machine learning techniques such as context-aware Recurrent Neural Networks (RNN), Long Short Term Memories (LSTM), regression analysis, image recognition-based monitoring systems, and state-of-the-art epidemiological early warning systems are capable of identifying the presence of late blight in farms, they are unable to help potato growers and supply chain managers with accurate early warnings and active monitoring, making them highly unreliable and unviable for scalable commercial applications in the global agricultural sector. A probable blight outbreak requires costly and labour-intensive treatments, which may result in bio-magnification and substantial capital loss if used very frequently over a period of time. A common farmer cannot rely on these technologies. However, a false negative in a blight risk assessment of a potato tuber can drastically reduce the production potential in a single cropping season, necessitating the urgent need for an accurate early warning and risk profiling approach. Based on our field research, interviews, data exploration and detailed experiments, we felt that an intelligent and robust agtech solution built with crucial capabilities such as predictive analytics, structural flexibility, precise analytics and early forecasting can dramatically improve the situation and provide new hopes for farming communities and agricultural researchers.

1. Proposed Solution
   1. Overview

To tackle the above-mentioned problems, we are introducing a new AI/ML and Graph-based solution for detecting the occurrence of a blight infection, model and monitor the spread pattern, predict upcoming blight infections and forecast potential future outbreaks of blight. This model consists of three components: A graph-based risk analysis system, an AI and Computer Vision-based diagnostic system, and an intelligent user reporting system for dynamically recalibrating the graph-based model. The decisions from each of these systems get combined through softmax and the outcome is fed back to the model to recalibrate the decision/learning parameters.



*Figure 1: Description and flowchart of Cilantro architecture*

* 1. Description of the Graph-Based Model

We start the process the segmenting any farming region into equally sized elementary zones and creating a graph structure for this entire area. Each of these elementary zones is treated as a node in this directed, weighted graph. In our solution, any arable land mass be represented as a directed, weighted graph of n nodes where each node represents a block..

Each block has the following properties:

1. Every block is connected to eight other blocks by direct land connection. (eight neighbors in the graph)
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.

It is evident from the above-mentioned properties that each block of land or wetland or waterbody can be treated as a node in a graph. 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 =

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 [ f1(t) f2(t) f3(t) f4(t) ]*T* 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) = Γ + c

Where,

*ϕ= constant that depends on the vector and disease*

*φ = constant that depends on the government policies and quarantine methods*

*Γ = 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 = .

Sn =

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 = P

Where,

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

The risk can be formulated as

R=

The value ηe-γt is called the Recovery Coefficient, which specifies the rate at which the effect of the disease dies off.

R, R,γ,μ,ζ ϵ (0,1]

k ϵ

α,β,κ ϵ (0,∞]

where,

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

*The constantsdepend on the particular disease.*

*n = the number of blocks or nodes.*

*μ = 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 is a sparse matrix. Thus, traditional sparse matrix representations are adequate in representing the graph. The Risk Factor 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.

* 1. Algorithm

The algorithm to simulate disease spread is:

**Inputs:**

t0 = initial time

tf := final time

d := time step

*j* := risk increase factor ∈ (0, 1]

*N* := {L:L is a block}

**Output**: Risk key value map *R\_table*

S**tart:**

1. *Adjacency\_matrix* := {}
2. *S\_table* := {}
3. *R\_table* := {}
4. **for** t **in** t0:tf **step** d:
5. Sum\_risk := 0
6. **for** L1 **in** N:
7. **for** L2 **in** N:
8. *Adjacency\_matrix*[L1][L2] := V(L1, L2)
9. **for** L **in** N:
10. *S\_table*[L] := S(L, t)
11. *R\_table*[L][t] := R(L, t)
12. Sum\_risk := Sum\_risk + *R\_table*[L][t]
13. Avg\_risk := Sum\_risk / n(N)
14. **for** L **in** N:
15. **if** *R\_table*[L][t] > Avg\_risk **then**
16. *R\_table*[L][t] := *R\_table*[L][t] \* (1 + *j*)
17. **return** *R\_table*

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

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.

This model will require large amounts of experimental data for validation and further optimization as well as tuning pH and Humidity have a major impact on the severity of the Blight outbreak. The severity can be approximated to be linear, logistic, or parabolic (GINI) according to the parameters considered.

1. Comparison with the existing state of the art

Existing state of the art technologies in the field of time series prediction typically use Recurrent Neural Network models, Regression, SVM, etc. These models are quite efficient in predicting temporal information one time step into the future. However, predicting deep into the future (>20 time steps) becomes erroneous, as the model predicts with some error and that same error is amplified into the future. Even predicting 6 time steps into the future causes error to skyrocket to about ~50% more than predicting 1 step into the future. This is particularly demonstrated in [https://doi.org/10.1016/j.neunet.2019.12.030], where LSTM and T-LSTM are used to demonstrate prediction capabilities of LSTM model. This is particularly crucial, since in case of Blight prediction, models need to predict more than 60 days into the future, for the prediction to have any economic viability. Blight treatment is costly and time consuming. It involves collecting and burning of infected tubers, which will be too late to do if the farmer is alerted on the ground only a day or even a week ahead, since by that time, the oospores and zoospores might have already bloomed on live hosts. Blight treatment needs significant capital investment, and an erroneous risk analysis can mean the difference between a profitable harvest and an irrecoverable loss.

Traditional machine learning models also fail when little or no data is provided to them in the training phase. The methodology of unsupervised and supervised learning is of use only when there is enough data to train a model accurately. In agriculture, properly filled, continuous data is often scarce. This can be because of a myriad of issues, like mishandling and mis-management at the government level, malpractice at the wholesaler level, erroneous observation at the grassroots level and so on. Cilantro, on the other hand can work with readily available data and is designed to fill in the inconsistencies in data, making it a software best suited for this application, precisely because it was built while keeping these aspects of agricultural data collection in mind.

* 1. Analysis of Predicting Blight data using RNN

Prediction of Blight using RNNs has mixed results. The inherent philosophy or algorithm for the application of any Machine Learning methodology rests on the assumption of availability of accurate and plentiful historical data. This condition is not always satisfied for crowdsourced agro-pathological data, especially in developing countries, where due to a combination of socio-economic factors and historical mistrust in government, farmers are often apprehensive to participate in schemes designed to collect data [33-35]. This makes model fitting and predicting difficult and erroneous. In the example dataset below, we see a similar thing happening. Many of the counties have no data for any blight outbreak. However, the areas with no data must also have blight spores, since the time within which the first outbreak is detected falls within the incubation period of oospores and zoospores of *Phytophthora Infestans.* In the data described in [23], there are four datapoints, in which, nearly 90% of the counties report non-availability of data. Those datapoints cannot be reliably backfilled, since the spores could already be present in those regions.

* 1. How CNN Fits into Cilantro

CNN models can accurately diagnose Blight by looking at leaf images. A separate CNN model deployed on edge devices can accurately generate crowdsourced dataset. Several commercial and open-source models and methodologies exist that can be deployed on Android/Swift platforms and can be made available to the common farmer. If we assume 10% of the farmers to be regular users, that can generate enough data for Cilantro to generate risk values for other regions, where there has been no reports. CNN models can similarly be used to generate reports and thus training data for RNN based models too. However, due to initial penetration issues for most Android and Apple applications in the commercial market, the data generated in the initial time can be sparse and inaccurate.

Reports from existing users 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. The sparse matrix representation of user reports is similar in shape and form to the risk matrix used to represent the risk of the next outbreak to enable the calculation and quantification of how far the prediction is from the actual data, and how the current scenario will change in the future.

1. Implementation of the Graph-Based Model

The entire graph-based software has been developed using Python.

To simulate the network of connected regions, we created a new python library called “regions.py”. This library includes the consummate simulation of the network predicated model that is required to be implemented.

The main class of this library is the “transmission\_graph” class, which represents a weighted directed graph data structure, wherein each node within the graph represents a separate region. The weight of the directed edge between any two nodes represents the probability of disease spread between these two nodes. The edge weights are computed based on the above-mentioned theoretical formula of the probability of disease spread between two regions. In this scenario, the graph was represented using an adjacency matrix. Because it is a fully connected graph with nodes connected to each other over some or all environmental constraints, an adjacency matrix representation is chosen over a sparse matrix representation.

1. Blight: An Overview

Blight is a kind of plant disease that is caused by bacterial or fungal infections that induces necrosis or cell death [31]. It attacks huge plant regions and destroys them swiftly. Tomatoes, potatoes, and apples, as well as many other kinds of ornamental plants, are among the commercially significant plants that are susceptible to one or more blights. *Phytophthora infestans* (Mont.) de Bary, a water mould, is responsible for late potato blight. Late blight may spread by the air, soil, and infected seed [30, 43] and has the potential to cause havoc anywhere potatoes are cultivated, but notably in the traditional potato-growing regions.

Requiring the administration of inorganic fungicides, the burning of crops and diseased foliage, the use of desiccant to completely kill off all plants in an affected area, and other hypersensitive techniques. Fungicides may effectively manage the disease in developed agricultural areas, but in many developing ones, small farmers find it difficult to afford them. There are more control options, but using them needs understanding the condition. Blight poses a serious threat to the food security of every country on Earth since the implementation of each of these strategies is intrinsically expensive.

The "Great Irish Famine" of 1843–1845 [28, 32], during which over a million people died because of famine and an additional 1.5 million fled to another nation, was mostly caused by late blight. The majority of Ireland was mostly dependent on a single potato type as its only food supply. Due to a wider variety of food available in other regions of Europe, starvation was avoided. The rural crops' lack of genetic diversity was a factor in the complaint's rigidity and its impact [28].

On stems and leaves of potatoes, recognisable big brown blotches can be observed, usually towards the leaf margin. The lesions spread swiftly and have the potential to kill the entire leaf. Despite better understanding of the illness, late blight still poses a significant global production challenge for potatoes. A disease called *Phytophthora infestans* that resembles a fungus and has secondary hosts including *Solanum nigrum, Datura* and *Stramonium* causes late blight in potatoes.

Together with the development of potatoes in Central and South America, the Late Blight fungus later spread to other parts of the world, primarily by contaminated seed tubers. Around 1830, the late blight of potatoes is believed to have spread from Central America to Europe and North America at approximately the same period. In 1840, late potato blight struck France, and in 1841, it wreaked havoc in Germany.

In 1943, reports of the illness Late Blight in Uttar Pradesh's plains were made. Major blight disease outbreaks took place in Punjab (the main potato-growing region of India) in 1985–1986; 1992–1993; and 2007–2008. Since 2010, reports of blight in potato and tomato crops towards the end of the summer in various locations of Alberta have appeared. This has led to the destruction or damage of several urban or homegrown plants, market garden commodities, and commercial potato harvests. The symptoms of late blight are frequently associated with increased susceptibility of plants to other diseases.

* 1. Symptoms and Disease cycle of Blight

The main sign of late blight is the brittleness and easy breakability of diseased potatoes. Over time, the leaves may deteriorate and become brittle. Pale to dark green patches that turn into blackish or brown lesions on leaves and stems are indicative of late blight in potatoes. These spots might start off tiny (a few millimeters across), seem wet or have chlorotic edges, but quickly spread until the entire leaf turns necrotic. Leaf tips and edges are typical sites where lesions start.

White sporulations are produced on the margins of lesions at abaxial (lower) leaf surfaces when the pathogen forms sporangia and sporangiophores at the surfaces of the diseased tissues in wet circumstances. Although the symptoms of late blight potato can be difficult to distinguish from those of other illnesses, the presence of spores and mycelia on the underside of leaves can add to the pathogenicity evidence. The pathogen that causes late blight is a minute, fungal-like creature. Its spores (sporangia) easily separate from the afflicted leaf and may move across great distances by wind.

Some infected potatoes may still have some tubers on the vine or even have brilliant green tubers, making them appear to be alive. This is because the potato maintains very little water in its tissues, despite the fact that the cells in the tubers are dying yet are still alive. The mould has the ability to penetrate potato tuber’s flesh and establish a colony there. Once it starts to infect the plant's tissues, the fungus changes.

As a result, secondary infections develop in the leaves and stems, which spread to other plants when contaminated tissue comes into touch with them. It inhibits the plant's ability to absorb nutrients, photosynthesize, and water. The interior tissues that contain the tuber's sucrose can be destroyed by the fungus, which can also cause it to decay and produce a characteristic odor.



*Figure 1 : Late blight of potato. Infected potato leaf [47]*

*Phytophthora infestans* (Mont.) de Bary reproduces both sexually and asexually. The pathogen is categorised as an obligatory parasite, meaning it needs a living host to exist.

Only under a microscope, phases of *P. infestans* (fungal threads or mycelium) life cycle can be observed. Contaminated tubers and infected soil are the principal sources of infection, and the fungus can survive in the soil as latent mycelium or in the fruiting stage. Affected field tubers may contain perennial mycelium that may be saved and utilised as seed the following year. Blight can develop at any stage, including blossoming [29].

For the development of illness, the pathogen prefers free moisture and cool to moderate temperatures of 50 to 70 F. Over watering creates the ideal environment for pathogen infection also thanks to free water from dew and rainfall. Young shoots are attacked in a very early stage because inoculum for the fungus is easily obtained from an already-infected primary-season crop. Late blight can spread from contaminated fruit and healthy tubers to stored crops (potatoes and tomatoes), including infected cull piles, uncompacted compost heaps, potato seedlings, and other *P. infestans* host plants. These crops include infected potatoes missed during harvest that remained unfrozen during the winter.

In the lack of an oospore stage, the pathogen lives on the infected tomato or potato tubers as a mycelium between potato crops. Given the thicker walls of sexual and asexual sporangia that are able to survive many seasons in the soil. Hence, the disease may get difficult to remove. Within 3–10 days of infection, sporangiophores emerge via the stomatal opening on the leaf surface during asexual reproduction. As there are more stomata present on the dorsal surface than the ventral surface, the sporulation is likewise more frequent on the dorsal side.

Conidia or zoosporangia grow at the tips of sporangiophores. These structures have a lemon-like form when seen under a microscope. Zoosporangia develop and then burst open, releasing all of their zoospores into the air, water, etc. These zoospores are said to travel up to 30 kilometres. These zoospores are motile for a short period of time or for several hours.This zoospore moves to a new host plant and uses a germ tube to pierce the epidermis of the leaves, stems, causing infection. The zoospores may begin a new life cycle after infection. In the absence of a suitable host, the zoospores die.

In sexual reproduction, one male form (antheridia) and the other female form (oogonia) of mating types (mycelia) A1 and A2 develop together. Up until 1984, only reports of

*P.infestans* A2 type came from Mexico and other Central American nations. Oogonium transforms into an oospore, which is a thick-walled spore, following mating and fertilisation. This oospore can endure harsh circumstances including droughts and freezing temperatures. This aids in the oospore's ability to endure the winter, dry spells, and lack of hosts.

This oospore development in *P. infestans* has not been seen outside of Mexico and Central America since there is no A2 mating type. Similar to zoosporangia reported in asexual reproduction, this oospore geminates. This oospore may begin a new life cycle upon the infection of the new host. *P. infestans* will finish its life cycle on a potato or tomato leaf in about five days under optimal circumstances.

* 1. Epidemiology and Management

The oospore can survive for a very long period of time in soil, but it can only persist within a live, compatible host.

* Lenticels and sores caused by tubers that get infected during or just before harvest.
* By diseased foliage, tubers may get contaminated.contaminated seed that can grow to produce an unusual potato.
* The main source of infection for the young plants may be cull heaps, which are leftover sick tubers from earlier harvests left on the field.
* It could originate from a nearby secondary host/collateral host that is near the potato crops.
* For the zoospore to penetrate, germinate, and develop, rain water and dew on the leaf surface for more than two hours are sufficient.
* Using tools and equipment to harvest an infected crop might spread spores.

If management measures are not implemented, potato late blight can spread exceptionally fast under ideal circumstances, with the ability to completely defoliate fields within three weeks after the first apparent infection. Fungicides should be applied promptly to prevent outbreaks, however the disease may be controlled once crops have been afflicted. Similar to potatoes, diseased tomato plants may get infected fast and die from the infection.

Numerous fungicides have been tested internationally against Late Blight in potatoes and tomatoes after more than 130 years after the introduction of the Bordeaux combination (copper sulphate, hydrated lime, and water) (first-generation Bordeaux mix through fourth-generation Mandipropamid and Azoxystrobin).

In order to avoid secondary inoculation and reduce the availability of primary inoculum (the initial waves of sporangia, or spores), late blight management strategies must be used (following waves of spores produced on plants). If it is possible that a few seed tubers are affected, using efficient fungicides will lessen the likelihood that late blight may affect the crops that follow. Some oomycete-specific fungicides appear to guard viable seed tubers during the seed-cutting process. Here, we find that oxathiapiprolin, when administered to soils where such potato plants are cultivated, is effective at preventing late blight on potato plants.

Management of disease requires timely fungicide treatments, cultural methods, and host resistance. All farmers, both large and small, including organic farmers, home gardeners, and other specialized producers, as well as governmental agencies, extension experts, and crop advisers, should embrace integrated management since blight is a disease that affects the entire community. Controlling infected sources, such as diseased tubers, volunteer plants, garbage piles, neighboring fields, and reinforcing growth after devastation by massive machinery, may help in the prevention of late blight.

A pre-emptive spray of mancozeb at 0.25 percent, followed by applications of cymoxanil + mancozeb or dimethomorph + mancozeb at 0.3 percent at the onset of the disease, and a second spray of mancozeb at 0.25 percent seven days after systemic fungicide application, can effectively manage severe late blight.

6. Experiments

To gauge the validity of the model, we try to predict the spread of the 2009 Late Blight Pandemic in the Eastern USA [23]. The Pandemic started synchronously in mid to late June over much of the northeastern USA. The pathway was via infected tomato transplants shipped to garden centers in large retail stores throughout the Northeast. Weather in the Northeast was relatively favorable to late blight development during June and July 2009. Prior to 2009 (and in most subsequent outbreaks), late blight occurred in production fields where fungicide applications were applied unevenly and rows or edges were missed. Extension personnel were surprised in 2009 because this outbreak was entirely different [23]. Homeowners and organic growers lost crops when fungicide was not applied soon enough, and because they lacked highly effective curative fungicide options. (Conventional commercial growers, who have more fungicide options, were more successful in delaying the epidemic and subsequent yield loss.) Other growers mistakenly thought their crops in high tunnels were protected [23].

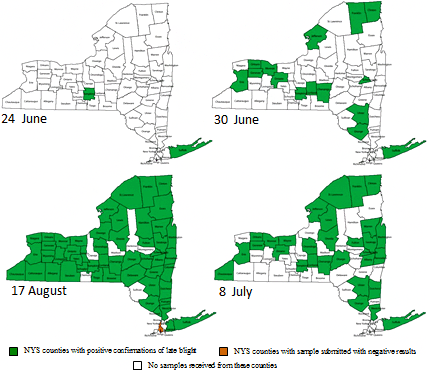
For this paper, we consider the spread of the pandemic between the dates June 24, 2009, and July 8, 2009, for 59 different counties in the state of New York, US. We tabulate the spread of the disease into a data frame with counties as rows and dates as columns. The cells have two values: 1 if samples 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. In essence, this can be modeled as a classification model. However, we are not implementing standard classification algorithms because of the gaps in the state of the art described in Section 1, which are shared by the standard ML algorithms.

* 1. 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 6.1, we have obtained disease spread information from the National Plant Diagnostic Network (NPDN) and have tabulated the data using tesseract and positional OCR.

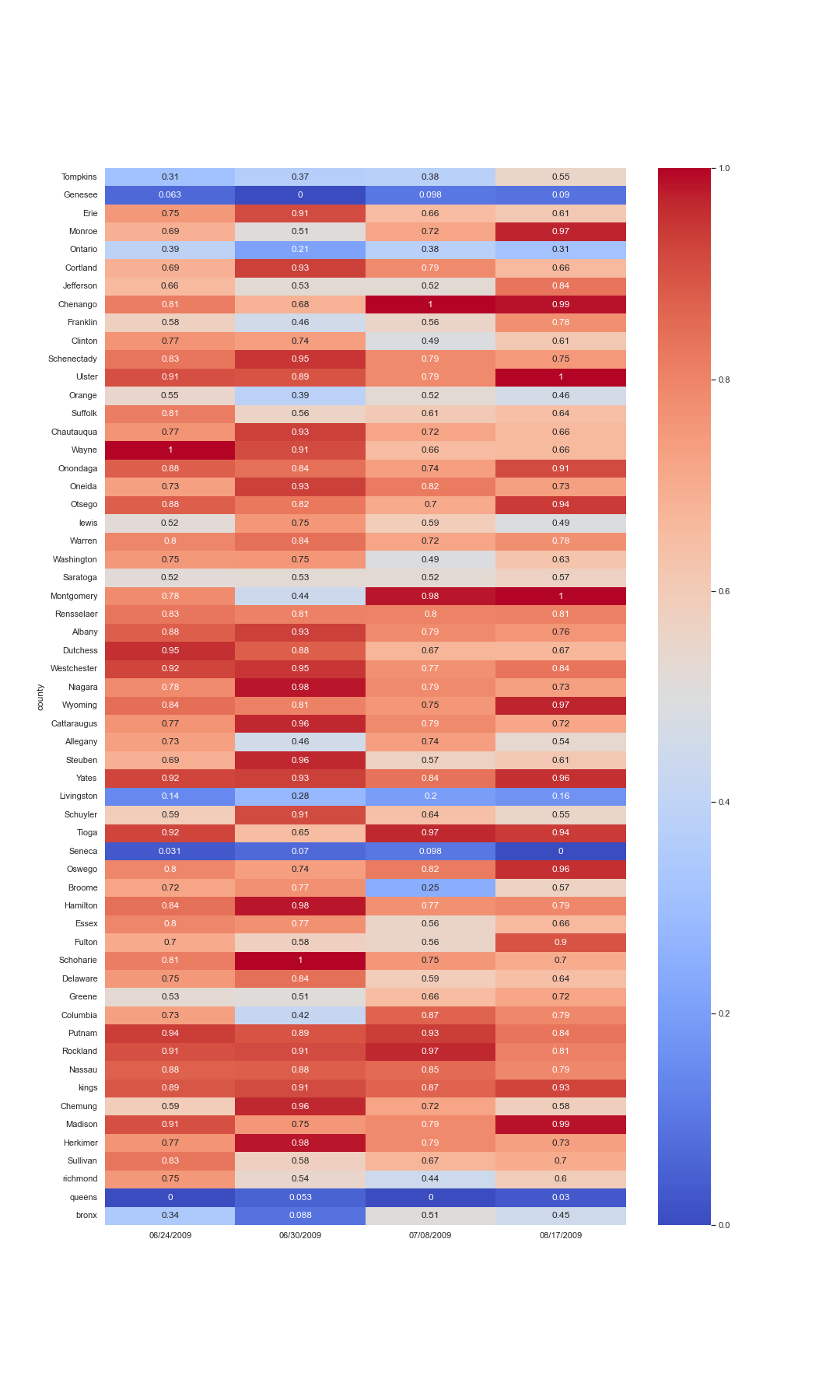


*Figure 3: Counties in New York which had reported late blight (in a garden center or in the field) during summer 2009. [23]*

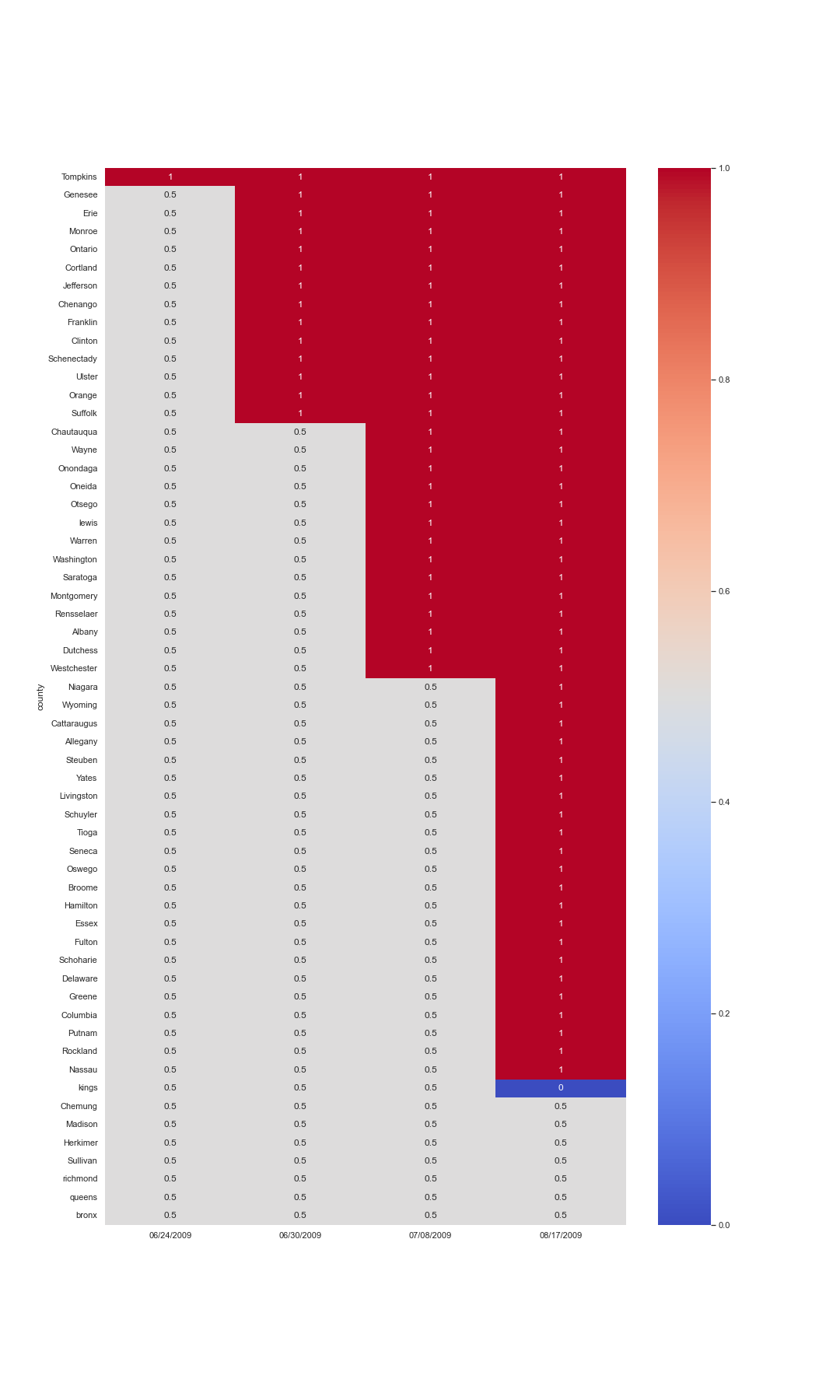
The dataset is converted into a graph-assisted spatiotemporal representation by first obtaining the latitudes and longitudes for the various counties represented on the map. Then the adjacency matrix is computed using the geographical distance between each county and the relative density of road interchanges within the county. In this study, road interchanges are considered, taking into account the dominance of road networks in the movement of crops in the US. 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 that takes into account the weighted incidence matrix and computes the real-world validation dataset as an array of two-dimensional matrices. The resultant dataset is fed into the model and the trends predicted by the model are obtained. (In progress)

1. Experimental Results and Analysis

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.

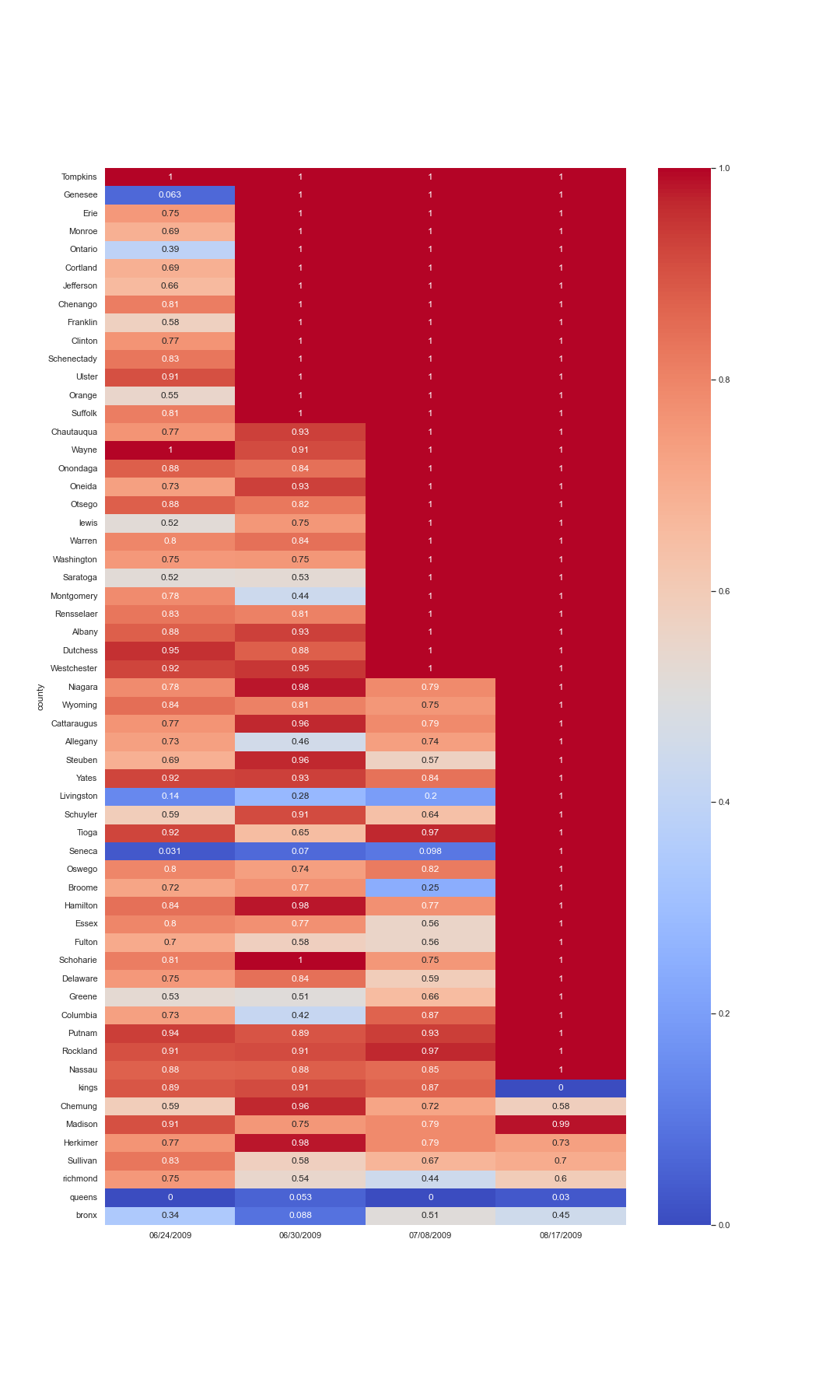


*Figure 1: The above diagram represent the heat map of the ground truth verses the predicted risk by the cilantro engine. Lighter colours shows the higher risk of disease and dark colour shows the lower risk of disease.*



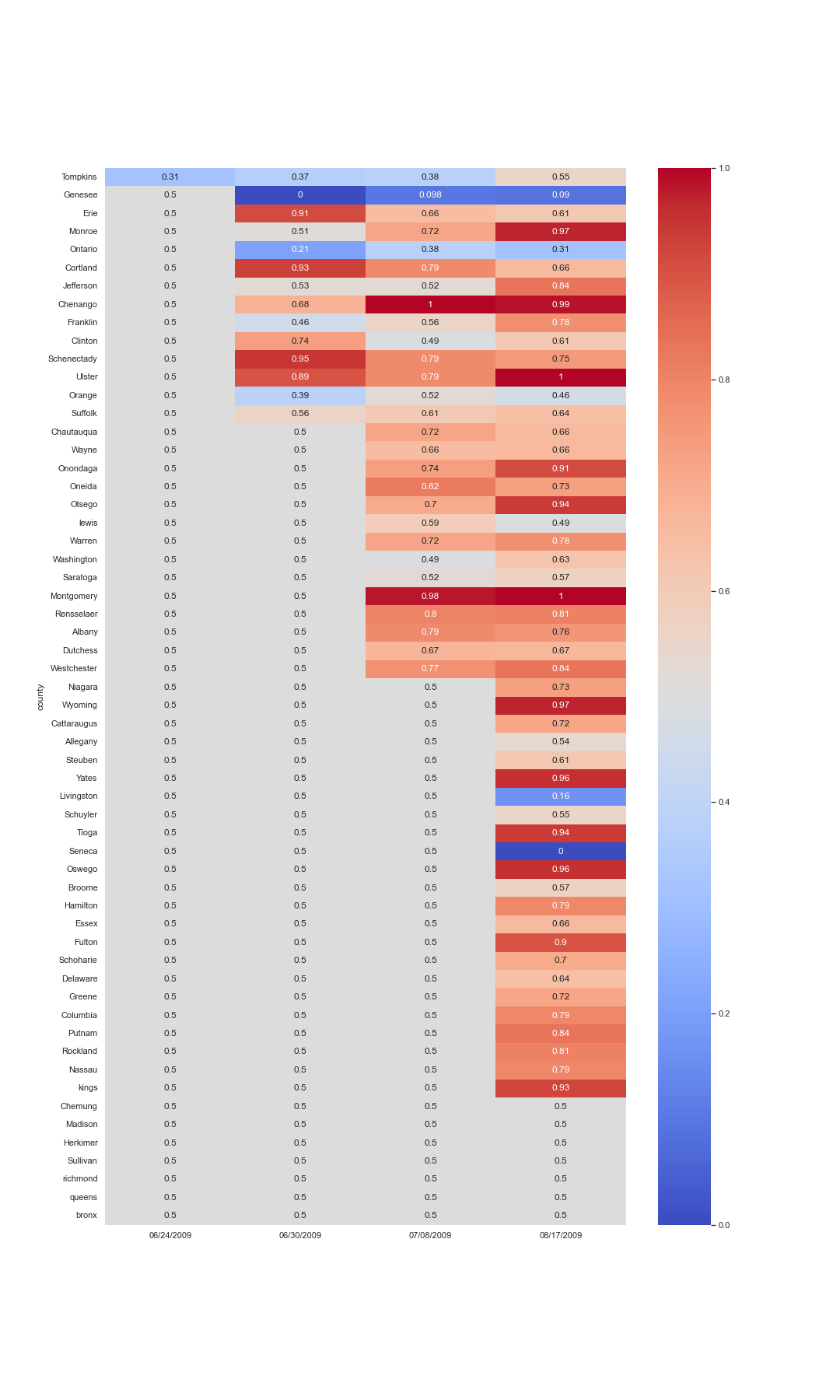
*Figure 2: In the diagram the risk is 0.5 when there is no report available. The risk is 0 when the report is false which means no blight is detected and it is 1 when blight is detected.*

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: This graph represents the risk profile of each county. Here are some risk value that the machine predicted 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.

When compared with RNNs, Cilantro has a substantial advantage in scaling and spatio-temporal information collection. RNNs inherently fail to predict blight risk very far into the future. We tried two data driven methodologies to predict Blight risk. In the first methodology, we took predictions of t+1 timestep from an LSTM model and used it to predict t+2, t+3, …t+60 timesteps into the future.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Timestep | t+1 | t+10 | t+30 | t+60 |
| Root Mean Squared Error | 0.01994 | 0.02395 | 0.03754 | 0.03898 |

The error increases to approximately 3 times its value when we predict t+60 into the future. A prediction just t+1 into the future will not be of practical or commercial use to the average farmer on the ground. A high blight risk profile presented only one day before the actual effects starts to show will not leave enough response time for the farmer to take necessary steps.

Cilantro also can project 60+ days days into the future without a huge drop in accuracy, making it a perfect candidate for commercial usage in scalable global early warning systems.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Timestep | t+1 | t+10 | t+30 | t+60 |
| Root Mean Squared Error | 0.01194 | 0.012 | 0.02634 | 0.02898 |

The second methodology that we tested involved training separate models for different counties, for different timesteps. To test t+60 timesteps in the future, we tested 60 separate GRU models. This approach is not scalable, since training an accurate RNN model takes enormous amounts of computational resources, time and data. Vanilla RNNs have a time complexity of O(n) where n is the length of the input sequence. Scaling it up for N nodes and assuming that 14% of the time steps involve retraining the model, that gives us a worst case time complexity of O(0.14 x t x N x n x T) or O(t.nNT.). In this case, generally, for a good prediction, n>>N. In total, we trained an ensemble of 3480 models, which took 14 hours on a Supercomputing Cray cluster with a theoretical peak performance of 377 TFLOPS, 128 compute nodes, a total of 4,686 CPU cores, 29 NVIDIA Kepler K20 GPUs, 4 NVIDIA Tesla P100 GPUs, 12 NVIDIA Tesla V100S GPUs, and 22.1 TB of memory. This amount of computing power is not cheap, making scaling this methodology difficult and unprofitable.

Cilantro has a worst case time complexity of O(t.N2) where N is the number of regions and t is the number of time-steps. Now,

n >> N => t.N.n >> t.N2 => O(t.N.n) >> O(t.N2)

Since Cilantro can be easily scaled up (by increasing the adjacency matrix), it provides a much more scalable and pragmatic prediction methodology useful for commercial deployment

1. Analysis of Results

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, Cortland, 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. Looking at the risk value of County Ontario, Orange, Livingston, Seneca,we can easily observe that the field data supported the risk value predicted by the engine. No blight outbreak was detected over the weeks.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.

In Chemung to Bronxcountiesno infection was detected. The ML engine produced a very low risk score which did not vary as much as the other counties.

The correlation between high-risk values and positive cases in respective counties proves the correctness of the model. To simplify the calculation, the function f(x) = 0.6667x is chosen as the decay value, while the vectors and atmospheric conditions are treated as independent random variables. However, further experiments with more field data and accurate empirical constants will be carried out to concretely validate the accuracy of *Cilantro*.

1. Future Work

As future work, we will concentrate on fine-tuning our model by further quantifying the disease severity scale. The quantification process will 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 and recalibration using remote sensing satellites.

1. Conclusion

In this article, we present *Cilantro*, a novel graph-based, scalable artificial intelligence algorithm that can predict the risk of a Blight outbreak 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.

1. References
2. 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]
3. 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]
4. 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]
5. 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]
6. 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]
7. Iglesias, I.; Escuredo, O.; Seijo, C.; Méndez, J. Phytophthorainfestans prediction for a potato crop. Am. J. Potato Res. 2010, 87, 32–40. [CrossRef]
8. 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
9. Kaukoranta, T. Impact of global warming on potato late blight: Risk, yield loss and control. Agric. Food Sci. 1996, 5, 311–327. [CrossRef]
10. Apel, H.; Paudyal, M.; Richter, O. Evaluation of treatment strategies of the late blight Phytophthorainfestans in Nepal by population dynamics modeling. Environ. Model. Softw. 2003, 18, 355–364. [CrossRef]
11. Aylor, D.E.; Fry, W.E.; Mayton, H.; Andrade-Piedra, J.L. Quantifying the rate of release and escape of Phytophthorainfestans sporangia from a potato canopy. Phytopathology 2001, 91, 1189–1196. [CrossRef] [PubMed]
12. Bruhn, J.; Fry, W. Analysis of potato late blight epidemiology by simulation modeling. Phytopathology 1981, 71, 612–616. [CrossRef]
13. Fall, M.; Van der Heyden, H.; Brodeur, L.; Leclerc, Y.; Moreau, G.; Carisse, O. Spatiotemporal variation in airborne sporangia of Phytophthorainfestans: Characterization and initiatives towards improving potato late blight risk estimation. Plant Pathol. 2015, 64, 178–190. [CrossRef]
14. 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]
15. Michaelides, S. A simulation model of the fungus Phytophthorainfestans (Mont) De Bary. Ecol. Model. 1985, 28, 121–137. [CrossRef]
16. 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.
17. 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 Phytophthorainfestans and implications for late blight epidemiology. Phytopathology 2015, 105, 449–459. [CrossRef]
18. 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]
19. 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]
20. 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]
21. Small, I.M.; Joseph, L.; Fry, W.E. Evaluation of the BlightPro decision support system for management of potato late blight using computer simulation and field validation. Phytopathology 2015, 105, 1545–1554. [CrossRef]
22. Van Oijen, M. Selection and use of a mathematical model to evaluate components of resistance to Phytophthorainfestans in potato. Neth. J. Plant Pathol. 1992, 98, 192–202. [CrossRef]
23. Narouei-Khandan, Hossein A., et al. "BLIGHTSIM: A new potato late blight model simulating the response of Phytophthorainfestans to diurnal temperature and humidity fluctuations in relation to climate change." *Pathogens* 9.8 (2020): 659.
24. Fry, W. E., et al. "The 2009 Late blight pandemic in Eastern USA." (2012).
25. 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. DOI: <https://doi.org/10.1016/j.enconman.2020.112766>
26. 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)
27. 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>
28. 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>
29. 28“Blight,” *Com.au*. [Online]. Available: https://www.yates.com.au/plants/problem-solver/diseases/blight/. [Accessed: 26-Sep-2021].
30. 29N. Alam, “Late Blight of potato *Phytophthorainfestans*. - name: azmeagri study,” *Nazmeagristudy.com*, 10-Jun-2021. [Online]. Available: www.nazmeagristudy.com/late-blight-of-potato-phytopthora-infestans/. [Accessed: 26-Sep-2021].
31. 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
32. 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]
33. 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)
34. M. Alkon and J. Urpelainen, “Trust in government and subsidy reform: Evidence from a survey of Indian farmers,” *Stud. Comp. Int. Dev.*, vol. 53, no. 4, pp. 449–476, 2018. DOI : <https://doi.org/10.1007/s12116-018-9266-1>
35. P. Chatterjee, “Agricultural reform in India: farmers versus the state,” *Lancet Planet. Health*, vol. 5, no. 4, pp. e187–e189, 2021. DOI: <https://doi.org/10.1016/S2542-5196(21)00060-7>
36. S. Gupta, “10 reasons why farmers are in distress,” *Times Of India*, 13-Dec-2018. [Online]. Available: https://timesofindia.indiatimes.com/india/10-reasons-why-farmers-are-in-distress/articleshow/67069814.cms. [Accessed: 12-Oct-2022].