Geoinformatica (2018) 22:363–381
https://doi.org/10.1007/s10707-017-0314-1

Spatio-temporal prediction of crop disease severity for agricultural emergency management based on recurrent neural networks

Wei Xu [1,2] · **Qili Wang** [1] · **Runyu Chen** [1]

Received: 31 January 2017 / Revised: 10 September 2017 / Accepted: 29 November 2017 /
Published online: 7 December 2017
© Springer Science+Business Media, LLC, part of Springer Nature 2017

Abstract As crop diseases bring huge losses every year in both developed and developing countries, determining how to precisely predict crop disease severity to facilitate agricultural emergency management is really a worldwide problem. Previous studies have introduced machine learning (ML) techniques into crop disease prediction and achieved better experimental results. However, the architectures of these ML models are unsuitable to model time series data. Moreover, the dependences among observations over time and across space have not been taken into account in model construction. By applying data-mining techniques to dynamic spatial panels of remote sensing data and considering features of bioclimatic, topographic and soil conditions as a supplement, we propose a novel crop disease prediction framework for agricultural emergency management based on ensemble learning techniques and spatio-temporal recurrent neural network (STRNN) which is an extension of recurrent neural network (RNN) in time and space. Empirical experiments are conducted on a specific dataset which is built based on reported cases of wheat yellow rust outbreaks in the Longnan city. Experimental results indicate that our proposed method outperforms all baseline models in crop disease severity prediction. The managerial implication of our work is that by applying the proposed methodology, some preparedness measures can be implemented in advance to prevent or mitigate the possible disasters according to predicted results. Notable economic and ecological benefits can be achieved by optimizing the frequency and timing of application of fungicide, pesticides and other preventative measures.

✉ Wei Xu
weixu@ruc.edu.cn

[1] School of Information, Renmin University of China, Beijing 100872, People's Republic of China

[2] Smart City Research Center, Renmin University of China, Beijing 100872, People's Republic of China

المنارة للاستشارات

Keywords Crop disease prediction · Deep learning · Recurrent neural network · Ensemble learning · Emergency management

1 Introduction

Crop diseases can cause huge losses in yield of crops. It is estimated that 16% of harvests in the eight most important crops is lost due to plant diseases that manifest during pre- and post-harvest treatment each year [30]. Agricultural emergency management measures reduce such losses by optimizing the frequency and timing of application of fungicide, pesticides and other preventative measures and ensure environmental and ecological safety by reducing excessive chemical application. A prediction model based on the relationship between environmental conditions before the time of management and crop disease severity can guide and facilitate management decisions. Thus, the explosive nature of the crop disease might be controlled if a reliable predictive model is developed. According to predicted results, some preparedness measures can be implemented in advance to prevent or mitigate the possible disasters. Like all predictors, crop disease prediction methods sometimes provide incorrect predictions. The frequency of incorrect predictions will definitely influence the rate of adoption and continue use of emergency solutions by potential users. In agricultural emergency management, in order to better configure fungicide, pesticides and other preventative measures, more accurate prediction models are needed. Thus, we undertook a case study on wheat yellow rust disease severity prediction by developing a novel deep learning model, spatio-temporal recurrent neural network (STRNN) and compared its experimental performance with the existing multivariate statistical approaches and machine learning techniques.

Wheat is grown on the most land area in the world among food crops. Caused by a kind of fungal pathogen called *Puccinia striiformis f.sp.tritici (Pst)*, yellow rust is one of the most damaging wheat diseases that occurs during the growing season in most wheat areas with moist and cool weather conditions [8, 16, 34, 43]. The occurrence of yellow rust, which produces leaf lesions that are yellow and tend to be grouped in patches, will lead to a severe impact on both the quality and yield of wheat [3]. The disease severity of the diseased leaf infected with *Pst* is usually estimated as a percentage figure. Environmental conditions, such as temperature, humidity, soil types and topography, are of vital important in the appearance and spread of the wheat yellow rust fungus. Due to high variability of the pathogen, almost all resistant varieties are also susceptible to infection so the most commonly used way to fight against yellow rust epidemics has been the use of fungicides. In order to reduce hazardous effects brought by excessive use of fungicides, early predictions of disease severity are required to avoid their unnecessary applications.

Previous studies have investigated the relationship between environmental conditions and crop disease severity through empirical experiments based on conventional multivariate regression analysis in various countries, including Ethiopia, Luxembourg, India, France and Belgium [14, 22, 24, 35, 40]. However, the problem of multi-collinearity has a side effect on prediction results since some environmental variables in the regression models are interrelated [23]. Moreover, contrast experiments indicate that the generalization ability of regression models for crop disease severity prediction is not good enough to be applied in practice [6]. Some pioneering papers have therefore, tried to introduce machine learning techniques into crop disease severity prediction for better results. Artificial neural networks (ANNs) based prediction models have been reported to outperform conventional multivariate regression

المنارة للاستشارات

techniques in crop disease prediction as ANNs can extract hidden subtle patterns from relevant multivariate data sources [12, 31, 36]. As machine learning (ML) methods keep on achieving outstanding results in various fields, attempts to apply advanced ML models, such as support vector machine (SVM) [6], bayesian networks [32] and random forest [29], in crop disease prediction continue to emerge and remarkably improve the prediction precision.

However, these ML-based architectures for disease severity prediction have some limitations in model settings. Firstly, most existing researches ignore the fact that environmental variables are complex multi-dimensional time series, like features of weather and climate data over a period of time. Thus, the high relevance of dependence between different points of the time series result in the inappropriateness of simply inputting data into fixed-size networks for data-mining. The dependence among observations over time should be paid more attention [1]. Secondly, the dependence among observations across space has not been taken into account in related ML models. More specifically, the disease severity and environmental conditions in neighboring areas will inevitably affect the disease infection in the near future, which should not be ignored in disease prediction. Thus, there is an urgent need of exploring a dedicated framework that exploit the latest advanced prediction techniques and give a more comprehensive consideration of the model construction for better understanding and prediction of the plant-disease-environment relationship.

In this paper, by applying data-mining techniques to dynamic spatial panels of remote sensing data and considering features of bioclimatic, topographic and soil conditions as a supplement, we propose a novel spatio-temporal crop disease prediction framework for agricultural emergency management based on ensemble learning techniques and STRNN which is an extension of recurrent neural networks (RNN) in time and space. Empirical experiments are conducted on a specific dataset which is built based on reported cases of wheat yellow rust outbreaks in the Longnan city for prediction precision analysis.

The rest of the paper is organized as follows. In Section 2, we review the related work on crop disease severity prediction and discuss the main differences between our study and previous researches. In Section 3, we introduce a framework that combines suites of environmental variables for crop disease severity prediction for agricultural emergency management. The design development of the proposed prediction model are implemented in Section 4. Section 5 introduces study area and data, evaluation metrics and experimental results. Finally, Section 6 concludes the paper and suggests further research directions.

2 Literature review

Crop disease prediction is of fundamental importance for the successful and efficient use of preventative agricultural emergency management measures to manage disease epidemics. Its specific goal is to predict the disease severity in a certain place at a certain future time by analyzing the relationship between environment conditions that may affect the appearance, multiplication and spread of the fungus and the severity of crop infection. In the past few decades, many research efforts have been devoted to investigating this plant-disease-environment relationship and improving the prediction accuracy to facilitate management decisions. These predictive analysis researches can be mainly divided into two categories through modeling methods: multivariate regression-based works and machine learning-based works.

Conventional multivariate regression analysis has been widely applied in previous studies. Coakley et al. [11] proposed an improved method to investigate interactions between

المنارة للاستشارات

meteorological variables and disease severity and quantified the relationship between climate and disease by developing statistical models. Te Beest et al. [37] identified the key factors determining the occurrence and severity of two kinds of disease epidemics on winter wheat and developed regression models with high-related factors to predict disease severity. Luo et al. [27] conducted a statistical analysis of relationship between land surface temperature (LST) and the occurrence of a crop disease and experimental results indicated that prediction models based on LST were effective. Landschoot et al. [25] firstly introduced ordinal regression technique into plant disease prediction to deal with ordinal scales. Their proposed method made it possible to predict whether certain severity thresholds will be exceeded as time goes by. The advantage of multivariate regression-based model is its interpretability. The impact of model variables on the output results can be well presented by significance and coefficient size. However, prediction accuracy is much more important than interpretability in disease severity prediction as the frequency of incorrect predictions will definitely influence the rate of adoption and continue use of emergency solutions by potential users.

With the development of artificial intelligence and big data analytics, it is a remarkable fact that prediction models based on data-mining and machine learning techniques are widespread and have achieved great success [13, 20]. A few pioneering papers have tried to apply ML models in crop disease severity prediction [7]. Wolf and Francl [12] was among the first researches that predicted crop disease with artificial neural network (ANN) technology. Comparative experiments in an outdoor environment showed that ANNs were more accurate than traditional statistical procedures. Kaundal et al. [22] selected 6 significant weather factors as dependent variables to develop weather-based disease prediction models based on SVM. Their case study showed that SVM outperform existing machine learning approaches and traditional regression analysis methods. Mehra et al. [24] compared prediction performance of three machine learning algorithms namely ANN, classification and regression trees and random forest (RF) on 431 disease cases. Results showed that RF was the most accurate algorithms with an accuracy rate of 93%. However, these existing ML methods ignore the fact that environmental variables like features of weather and climate data over a period of time are complex multi-dimensional time series. The high relevance of dependence between different points of the time series contain valuable hidden information.

This paper develops a novel spatio-temporal ensemble learning framework for crop disease prediction based on STRNN which is an extension of RNN in time and space. Allowing cyclical connections in the network, a recurrent neural network can learn from all previous inputs to obtain output. RNNs have done well on many sequence learning problems by virtue of their internal memory [9, 10, 21, 26, 39]. There is a universal approximation theory for RNNs. An RNN can achieve arbitrary accuracy in approximating measurable sequence-to-sequence mapping, as soon as it has sufficient number of hidden units [18]. Along with popular agreement that RNNs are well suited to process, classify and predict time series, they have begun to be applied in more fields, where time series data are a common data type [2, 33, 41].

To sum up, the main contributions of this research are fourfold. Firstly, we design a novel data-mining framework for spatio-temporal prediction of crop disease severity for agricultural emergency management. Secondly, a new deep learning technique, RNN, which is more suitable for sequence learning modeling, is introduced in crop disease prediction. Thirdly, we develop a STRNN-based ensemble learning model that considers both spatial and temporal dependency in model settings for more effective spatio-temporal mining of possible disease outbreaks. At last, an empirical analysis is performed to test the effectiveness of the proposed

المنارة للاستشارات

method. To the best of our knowledge and belief, this is the first successful research work on applying a deep learning technique to analyze and predict crop disease severity.

3 A Spatio-temporal prediction framework

The appearance, multiplication and spread of the crop disease fungus will be dynamically influenced by local and neighboring environmental conditions. In order to better understand and predict the plant-disease-environment relationship, we propose a dedicated spatio-temporal framework for wheat yellow rust severity prediction. For the proposed framework, by summarizing environmental variables used in previous studies, we take bioclimatic, remote sensing, topographic and soil data into account for prediction of crop disease severity. In particular, we develop a STRNN-based ensemble learning model for more effective spatial mining of possible crop disease outbreaks. The proposed prediction framework consists of four main processes, namely data discovery, data organization, data modeling, and spatio-temporal prediction, which is outlined in Fig. 1.

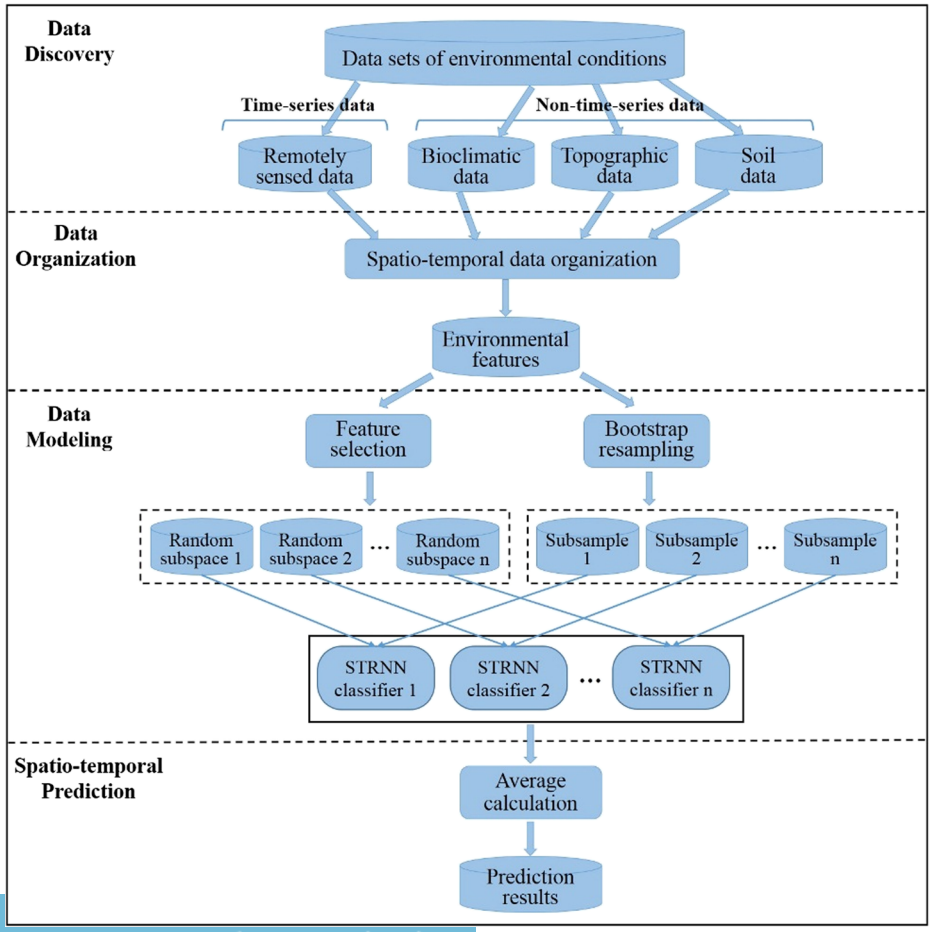

Fig. 1 A framework for wheat yellow rust severity prediction

المنارة للاستشارات

3.1 Data discovery

During the first step of the proposed framework, data discovery, data sets of environmental variables required for crop disease prediction are searched for and collected. The pathogenesis and infection characteristics of wheat yellow rust are explored by consulting relevant literatures for domain knowledge. According to previous studies, climatic, topographic and soil conditions are associated with the possibility of wheat yellow rust outbreaks, which in a way demonstrates the feasibility of predicting future disease severity based on various environmental variables [8, 40]. For example, cool and wet environmental conditions during the growing season are known to favor *Pst* infection and spread which inspires us to take temperature and humidity time series as independent variables of the prediction model to improve its predictive capabilities. In order to better exploit the implicit relationship of those environmental conditions leading to disease occurrence, more comprehensive data sources are needed. We discovery environmental data sets from four aspects: bioclimatic, remote sensing, topographic and soil variables. The selected data sets consist of time-series data which change over time and non-time-series data which remain stable for a period of time.

3.2 Data organization

Data organization, in broad terms, refer to the approach of organizing and classifying data sets to make them more specific and useful [4, 38, 42]. In the field of spatio-temporal prediction, spatial data fusion should be conducted on multiple spatial data sources firstly. Data fusion is the process of integrating multiple knowledge and data representing the same real-world object into an useful, accurate and consistent representation, which facilitates further data analysis and mining [17, 28]. The study areas are divided into observation spots of the same size with unique coordinates, which are basic units for data organization, modeling and prediction. The spatial resolution of these observation spots in our prediction tasks is the minimum spatial resolution of all data sources. Then the spatial uniform data should be arranged in chronological order to prepare for the time-series data mining. Data preprocessing, including data cleaning, imputation, transformation and normalization, is also done in this step. The phrase "garbage in, garbage out" clearly illustrates the importance of data processing in data mining since knowledge discovery will be more difficult if there is too much noise in input data during training phase. Afterwards, suits of environmental features, including remote sensing, bioclimatic, topographic and soil features, are extracted from the corresponding data sets for the data modeling.

3.3 Data modeling

As described above, four kinds of environmental data are combined for crop disease severity prediction. The environmental features of spot (x,y) at time t are organized as $S_{x,\,y,\,t} = (X_{x,\,y,\,t}, M_{x,\,y}) = (X_{x,\,y,\,t}, \{J_{x,\,y}, K_{x,\,y}, L_{x,\,y}\})$, where $X_{x,\,y,\,t}$ represents the $N_X \times 1$ vector of remote sensing features at time t, $J_{x,\,y}$ represents the $N_J \times 1$ vector of the bioclimatic features, $K_{x,\,y}$represents the $N_K \times 1$ vector of the topographic features and $L_{x,\,y}$ represents the $N_L \times 1$ vector of the soil features on the same day. N_X, N_J, N_K and N_L represent the corresponding number of features. $M_{x,\,y}$ represents a set of non-time-series features. As $J_{x,\,y}$, $K_{x,\,y}$ and $L_{x,\,y}$ remain stable for a long period of time, we do not consider their short-term changes. The original prediction problem of

المنارة للاستشارات

estimating crop disease severity of spot (x,y) at time t given the vector S_t of environmental data can be presented by the formula below:

$$DS_{x,y,t} = f\left(S_{x,y,t-1}\right) = f\left(X_{x,y,t-1}, M_{x,y}\right) = f\left(X_{x,y,t-1}, \left\{J_{x,y}, K_{x,y}, L_{x,y}\right\}\right) \tag{1}$$

where $DS_{x,\ y,\ t}$ denotes the predicted disease severity at time t. The dependences among observations over time and across space have not been considered in this simple formula. By considering both temporal and spatial dependency of independent and dependent variables in model settings, a complete spatio-temporal prediction problem can be presented as follows:

$$DS_{x,y,t} = f\begin{pmatrix} X_{x,y,t-1}, X_{x-1,y,t-1}, X_{x+1,y,t-1}, X_{x,y-1,t-1}, X_{x,y+1,t-1}, \ldots, X_{x,y,t-2}, X_{x-1,y,t-2}, \\ X_{x+1,y,t-2}, X_{x,y-1,t-2}, X_{x,y+1,t-2}, \ldots, M_{x,y}, M_{x-1,y}, M_{x+1,y}, M_{x,y-1}, M_{x,y+1}, \ldots, \\ DS_{x,y,t-1}, DS_{x-1,y,t-1}, DS_{x+1,y,t-1}, DS_{x,y-1,t-1}, DS_{x,y+1,t-1}, \ldots, \\ DS_{x,y,t-2}, DS_{x-1,y,t-2}, DS_{x+1,y,t-2}, DS_{x,y-1,t-2}, DS_{x,y+1,t-2}, \ldots \end{pmatrix} \tag{2}$$

As discussed above, crop disease severity prediction can be regarded as a special type of sequence learning problem. Machine learning methods aim for more accurate prediction results by investigating and exploring hidden pathogenesis patterns relying on multiple environmental variable sequences. Given that RNNs achieve outstanding performance in dealing with synthetic tasks that require long-range memory, in this paper, we design a spatio-temporal RNN model to solve the disease severity prediction problem for agricultural emergency management. In this model, environmental data are mapped to real value vectors in the input layer after normalization, which guarantees that the proposed model can deal with different features uniformly. Moreover, in order to improve prediction accuracy and enhance procedural bias, ensemble techniques including bootstrap aggregating and random subspace methods are applied to construct several base STRNN classifiers for ensemble learning. The design development of the proposed deep learning based ensemble learning model will be introduced in Section 4.

3.4 Spatio-temporal prediction

Our work aims to investigate hidden patterns of the crop disease occurrence relying on multi-dimensional environmental data and predict the disease severity to facilitate agricultural emergency management. For our case dataset, each case is labeled with a disease severity which is assessed according to the Rules for Monitoring and Forecast of the Wheat Yellow Rust (National Standard of the People's Republic China, GB/T 15795–2011). We estimate the disease severity as 1%, 5%, 10%, 20%, 40%, 60%, 80% or 100%.

The ground truths of the crop disease prediction tasks are represented by these labels. The disease severity of contrast cases where there is no disease occurrence is set to 0%. A balanced label distribution is maintained to avoid serious learning biases [5, 15]. After the training process, a STRNN model is established for the future spatio-temporal disease severity prediction. The final predictive model is established by average calculation of several base STRNN classifiers to predict disease severity from the test set. Then, we compare the prediction accuracy with some baseline models on the same evaluation dataset to assess the effectiveness of our proposed method.

المنارة للاستشارات

4 The computational details

In this study, environmental variables that reflect the growth environmental conditions of wheat consist of time-series variables and non-time-series variables. Changes in environment conditions over time may lead to changes in disease occurrence and disease severity, which inspire us to investigate time-series environmental variables using machine learning techniques suitable for time-series modeling. Moreover, spatial dependency is also considered during model development. We aim to develop an efficient spatio-temporal deep learning model based on the extension of RNN for prediction of crop disease severity. The design development of the proposed STRNN model is introduced in this section.

As shown in Section 3.3, formula (1) indicates a non-spatial functional relationship between independent and dependent variables in estimating crop disease severity. Since the disease severity and environmental conditions in neighboring areas will inevitably affect the disease infection in the near future, it is necessary to extend the basic non-spatial model with spatial interaction effects. The first one is endogenous interaction effect, where the dependent variable of a particular spot depends on the dependent variables of other spots. In other words, the disease severity of an observation spot will be influenced by the disease severity of other spots. We assume that the influence domain is a square and the length of its sides is *2D*, so the total number of spots in the influence domain is $(2D+1)^2-1$. A sketch map of spatial influence domain is outlined in Fig. 2. The central unit in the figure represents the target spot. Then the functional relationship is updated as follows:

$$\begin{aligned} DS_{x,y,t} &= f\left(S_{x,y,t-1}, DS_{x,y,t-1}, \sum_{(x',y')\in ID} DS_{x',y',t-1}\right) \\ &= f\left(S_{x,y,t-1}, DS_{x,y,t-1}, DS_{x-1,y,t-1}, DS_{x+1,y,t-1}, \ldots\right) \end{aligned} \tag{3}$$

where *ID* refers to the influence domain. Upper-case $\sum$ is used as a symbol for enumeration operator instead of summation operator here.

The second spatial interaction effect is exogenous interaction effect, where the dependent variable of a particular spot depends on the independent variables of other spots. It means that the disease severity of a spot will be influenced by environmental conditions of other

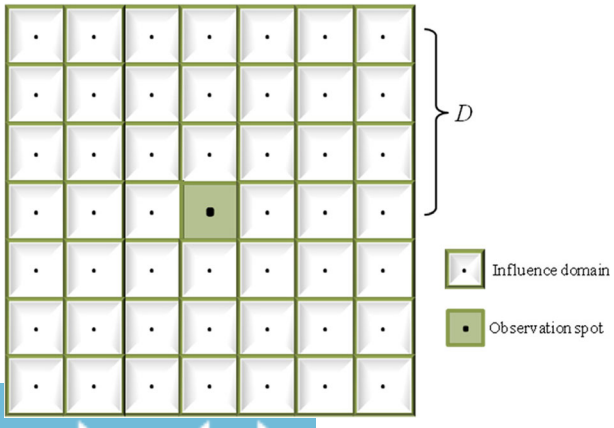

Fig. 2 A sketch map of spatial influence domain

neighboring spots. After considering these two spatial interaction effects, the complete form of spatial model can be presented.

$$DS_{x,y,t} = f\left(S_{x,y,t-1}, DS_{x,y,t-1}, \sum_{(x',y')\in ID} S_{x',y',t-1} \sum_{(x',y')\in ID} DS_{x',y',t-1}\right) \tag{4}$$

The above model only includes variables with serially lagged value $t-1$. Next, the dependences among observations over time are also considered. We extend the modeling data sets to dynamic spatial panels and use variable T to indicate the length of the panel. Formula (5) is the final function form of the spatio-temporal prediction model. The upper-case $\sum$ is still used as a symbol for enumeration operator here.

$$\begin{aligned} DS_{x,y,t} &= f\left(S_{x,y,t-1}, DS_{x,y,t-1}, \sum_{t'\in[t-T,t-1]} S_{x,y,t'} \sum_{t'\in[t-T,t-1]} DS_{x,y,t'} \sum_{\substack{t'\in[t-T,t-1]\\(x',y')\in ID}} S_{x',y',t'} \sum_{\substack{t'\in[t-T,t-1]\\(x',y')\in ID}} DS_{x',y',t'}\right) \\ &= f\left(\sum_{t'\in[t-T,t-1]} \left(S_{x,y,t'}, DS_{x,y,t'}, \sum_{(x',y')\in ID} S_{x',y',t'} \sum_{(x',y')\in ID} DS_{x',y',t'}\right)\right) \end{aligned} \tag{5}$$

All variables in the spatial model with the same time stamp are organized as input vectors for recurrent modeling using RNN. The total number of features N is $(2D+1)^2 \times (N_X + N_J + N_K + N_L + 1)$. Then we employ a double disturbance strategy to keep more differences in an ensemble system to improve procedural bias through disturbance of both the samples and the feature spaces. Bootstrap method is applied to randomly sample d features to construct a random subspace ($d < N$). By repeating this feature selection step n times, n random subspaces are obtained. Similarly, n subsamples are generated using bootstrap resampling on the training set. Each STRNN model is trained based on the combination of a random subspace and a subsample. The final predictive model is established by average calculation of n base STRNN classifiers to predict disease severity from the test set. In the proposed framework, spatio-temporal deep learning algorithm is employed to enhance disease severity prediction capabilities and ensemble learning techniques are used to avoid over-fitting and improve the overall performance.

Backpropagation is performed to compute the gradients from the output to the input using the chain rule. For base STRNN model training, stochastic gradient descent is applied using backpropagation through time (BPTT) algorithm according to the characteristics of the proposed model structure. The training process continues until all weight matrices reach convergence. After the network has been trained and established, the output is defined as the estimate of future disease severity given the relevant dynamic spatial panels of data. The unfolded form of a single STRNN model is shown in Fig. 3.

5 The empirical analysis

5.1 Data description

Wheat is grown on the most land area in the world among food crops and China is the world's largest producer of wheat. The Longnan city, an important pathogen of wheat yellow rust, is

المنارة للاستشارات

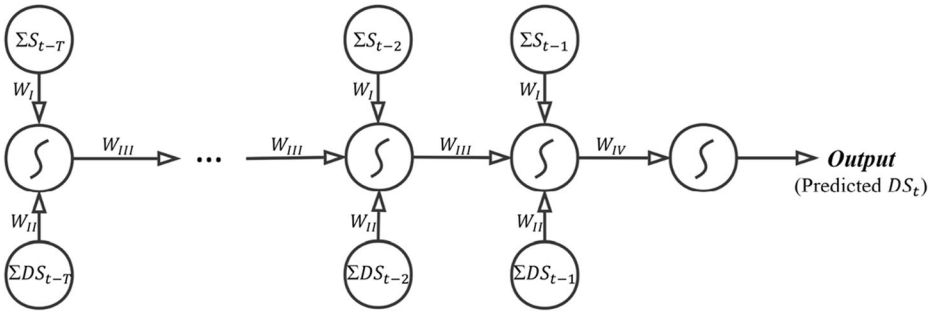

Fig. 3 Unfolding a single STRNN model as a deep network

located in the southeastern parts of the Gansu province, China. Elevation in Longnan ranges from 500 m to 4187 m above sea level. Pathogenicity variation of wheat yellow rust is the main cause of disease loss caused by "loss" of rust resistance of wheat cultivars. Longnan is one of the wheat yellow rust easily mutated areas, where resistant varieties will lose resistance in just a few years, so it is of great importance to take early control of yellow rust through accurate spatio-temporal prediction of the disease severity for emergency management in the near future (Fig. 4).

By summarizing environmental variables used in previous studies, a suite of topographic, soil, bioclimatic and remote sensing variables are selected to model wheat yellow rust severity in Longnan. MODIS normalized difference vegetation index (NDVI) and leaf area index (LAI) are used to reveal chlorophyll content and wheat health. Temperature is known to be a key meteorological factor that influences the distribution of *Pst* and MODIS land surface temperature (LST) is used to represent the field temperature. Gross primary productivity (GPP) and net primary productivity (NPP) can reflect the wheat growth situation over a period of time. By using the MODIS reprojection tool web interface, all selected MODIS layers were downloaded from the Land Processes Distributed Active Archive Center of the National Aeronautics Space Agency (NASA LPDAAC). The widely known 19 bioclimatic moisture- and temperature-related variables [19] (http://www.worldclim.org/) are included in the proposed model to describe the bioclimatic conditions. Compound

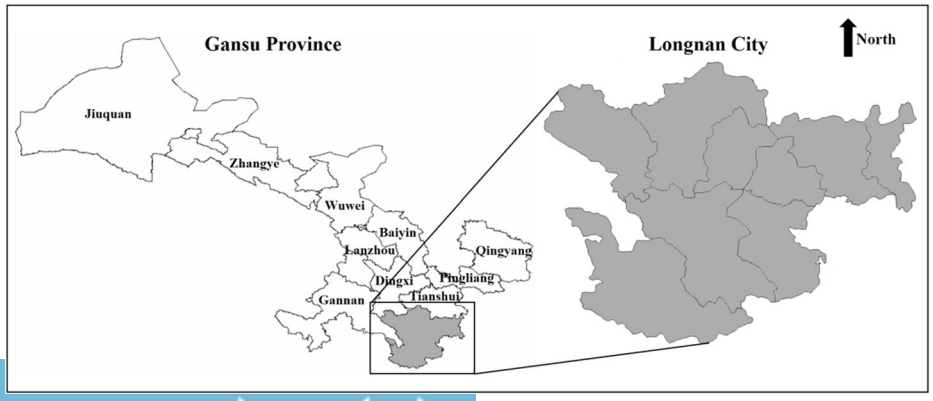

Fig. 4 Map of the study area

topographic index (CTI) and elevation, derived from the Digital Elevation Model of the Shuttle Radar Topography Mission (SRTM DEM), are selected to represent topographic heterogeneity. All soil variables are obtained from the ISRIC-WISE database. We aim to mining time series of these environmental variables based on spatial information to carry out spatio-temporal disease severity prediction. According to news reports and historical records about epidemic situation of wheat yellow rust during 2010 to 2016, we collected 187 disease severity cases randomly in Longnan. The specific selections of environmental variables are listed in Table 1.

5.2 Evaluation metrics

The popular 5-fold cross-validation method, which can make full use of every available case and generally achieve a good trade-off between model over-fitting and under-fitting, is employed for evaluation of the proposed model. Thus, the dataset is randomly partitioned into five equally sized slices. For each of the 5 folds, one slice is selected for testing and the remaining 4 slices are used for training the model. In this way, we can test the performance of the model on every available case without having used it in the prior training phase.

Table 1 Environmental variables

Variable types	Variable name	Data source
Remote sensing	Normalized difference vegetation index (NDVI)	NASA LPDAAC
	Land surface temperature (LST)	
	Leaf area index (LAI)	
	Gross Primary Productivity (GPP)	
	Net Primary Productivity (NPP)	
Bioclimatic	Max temperature of warmest month	WorldClim
	Min temperature of coldest month	
	Temperature seasonality (standard deviation)	
	Isothermality (diurnal range / annual range)	
	Annual mean temperature	
	Mean diurnal range	
	Annual precipitation	
	Mean temperature of coldest quarter	
	Mean temperature of warmest quarter	
	Mean temperature of driest quarter	
	Mean temperature of wettest quarter	
	Temperature annual range	
	Precipitation of wettest quarter	
	Precipitation of driest month	
	Precipitation of wettest month	
	Precipitation seasonality (coefficient of variation)	
	Precipitation of coldest quarter	
	Precipitation of warmest quarter	
	Precipitation of driest quarter	
Topographic	Compound topographic index(CTI)	SRTM DEM
	Elevation	
Soil	Total nitrogen content(top 20 cm soil horizon)	ISRIC-WISE
	Clay mass(top 20 cm soil horizon)	
	Water holding capacity(top 20 cm soil horizon)	
	Soil types	

المنارة للاستشارات

The test results are evaluated by three commonly-used performance measures, mean absolute error (MAE), mean absolute percentage error (MAPE) and root-mean-square error (RMSE), which are defined as follows.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{6}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{|\hat{y}_i - y_i|}{y_i}\right| \times 100\% \tag{7}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{8}$$

where $\hat{y}_i$ denotes the predicted disease severity value and y_i denotes the observed disease severity value, respectively. The lower these evaluation metrics, the more accurate the predictions.

5.3 Experimental results

The experimental results under different parameter settings are shown in this section. The impacts of the side length of the influence domain ($2D$) and the length of the dynamic spatial panel (T) on the proposed STRNN-based ensemble learning (STRNN-EL) model are investigated. The meaning of these two parameters has been introduced in Section 4. They are two key parameters in the spatio-temporal prediction framework. In the model training, the learning rate is set to 0.05 and the regularization parameter is set to 1×10^{-7}. According to the experimental results, optimal parameter settings are determined.

Firstly, as an attempt, half of the side length of the influence domain (D) is set to 3. Then, the values of the evaluation metrics MAE, MAPE and RMSE of the proposed prediction model are calculated in the cases of $T = 1, 2, 3, 4, 5, 6, 7, 8, 9$. The results are shown in Fig. 5. As T increases, the MAPE falls rapidly at the beginning and then grows gently. The MAE and RMSE curves have a similar trend. When the length of the dynamic spatial panel equals 5, the proposed model achieve the best experimental performances, with a MAPE value of 22.03%, a MAE value of 0.0424 and a RMSE value of 0.0547. It is indicated that the environmental factors a few days before the forecast date do have an impact on wheat yellow rust severity which can be recognized by our prediction model, as its recurrent structure remembers the former input of time-series data over a fixed time window. However, dramatic changes in these performance measures occur when the length of the dynamic panel gets longer, the time window becomes larger, and the noise in the time-series data begins to influence the prediction accuracy of the model. Secondly, according to the analysis of the results of the first experiment, the length of the dynamic spatial panel (T) is set to 5. Then, we investigate the impact of half of side length of the influence domain (D) on the prediction model. The results are shown in Fig. 6. When D is lower than 4, the three performance measures falls rapidly as D increases. But as D gets higher than 4, the MAPE begins to grow and the other two measures increase slowly at first and then remain stable. These results indicate that the increase of spatial environmental information significantly improves prediction capabilities of the STRNN-EL model, but there may exist over-fitting when the area of the influence domain is very large.

المنارة للاستشارات

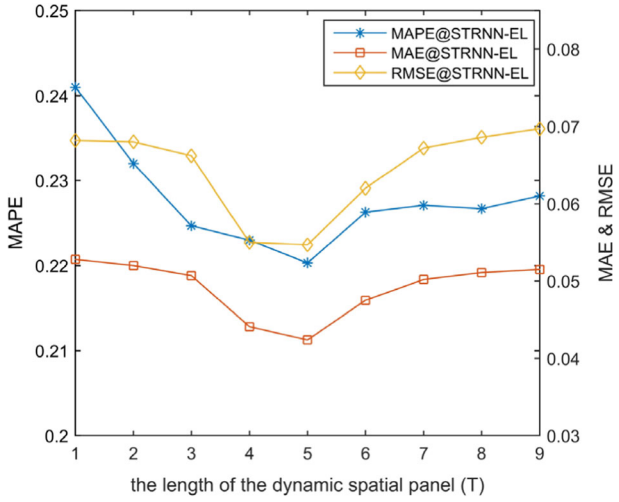

Fig. 5 Impact of the length of the panel

5.4 Comparison

As introduced in Section 2, previous studies have mainly applied multiple linear regression (MLR) and logistic regression (LR) to investigate interactions between meteorological variables and disease severity and predict future disease occurrences by developing statistical models. Recently, prediction models based on machine learning techniques have achieved better results than traditional statistical methods in crop disease prediction. Given the lack of consensus on the best prediction method, we incorporated several popular classifiers that have been applied in previous studies as baseline models in comparative experiments. The selected classifiers are as follows: MLR, LR, artificial neural network (ANN), support vector machine (SVM) and random forest (RF). In this section, two groups of comparative experiments were carried out on the same evaluation dataset to evaluate the effectiveness

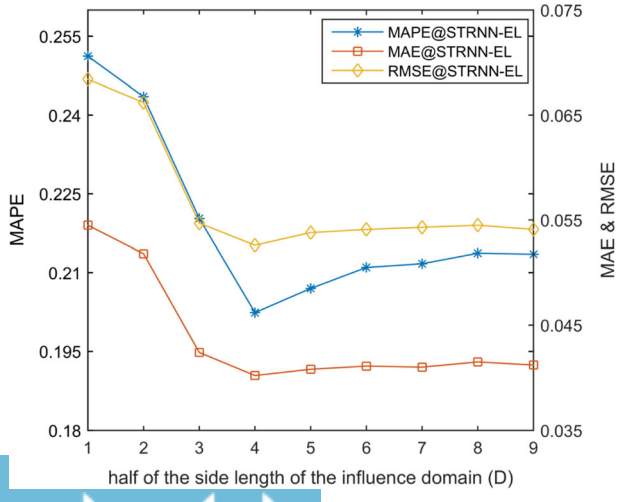

Fig. 6 Impact of the side length of the ID

of the proposed spatio-temporal crop disease prediction framework, each of which was designed to assess the utility of one facet of the framework. Experiment group 1 (presented in Section 5.4.1) evaluates STRNN model in comparison with five baseline models in predicting wheat yellow rust severity with non-spatial environmental time-series variables. Experiment group 2 (presented in Section 5.4.2) evaluates the utility of considering spatial dependency as supplementary information in the prediction model. The efficacy of using ensemble learning techniques is also tested in this subsection.

5.4.1 Evaluating time-series modeling

STRNN is served as base classifier in the proposed spatio-temporal prediction framework as RNNs have done well on many synthetic sequence learning tasks by virtue of their internal memory [10, 21, 39]. Previous regression-based models mainly included environmental time-series variables. We want to investigate which model can better predict disease severity with only time-series variables. Statistical methods (MLR, LR), general machine learning models (ANN, SVM and RF) or RNN? We incorporated paired t-test to test the significance of the performance improvements of STRNN model over the best baseline. A *p*-Value is calculated to show the statistical significance of an improvement. If the p-Value is lower than 0.05, the difference between two evaluation results can be considered statistically significant. The lower the p-Value, the more significant the improvements. For non-time-series modeling, those time-series variables are transformed into one-dimension vectors. For example, if the number of time-series variables is N_x and the length of time window in the recurrent modeling is T, then the total number of variables is $N_X \times T$ for other models. The evaluation results are shown Table 2.

As can be seen from Table 2, STRNN model outperforms the five baseline models on all three performance measures. It improves over the best baseline by 30.20% on MAE, 23.31% on MAPE and 44.52% on RMSE, respectively. Moreover, these improvements are all statistically significant. The results indicate that RNN-based model is better suited for crop disease severity prediction than traditional statistical methods and general machine learning techniques. The comparative experiments also provide empirical support for the conclusion of Kaundal et al. [22] that machine learning models perform better than traditional statistical algorithms in the field of crop disease prediction.

5.4.2 Evaluating the effectiveness of including spatial information

As introduced earlier, the dependence among observations across space has not been taken into account in related ML models. As a matter of fact, the disease severity and environmental

Table 2 Performances of various methods using time-series variables

Models	MAE	MAPE	RMSE
MLR	0.1431	61.87%	0.2009
LR	0.1201	49.41%	0.1818
ANN	0.1147	46.36%	0.1706
SVM	0.0990	42.91%	0.1620
RF	0.0937	39.72%	0.1543
STRNN	0.0654	30.46%	0.0856
Improvements over the best baseline	30.20%**	23.31%**	44.52%*

* p-Value is lower than 0.05 ** p-Value is lower than 0.01

المنارة للاستشارات

conditions in neighboring areas will inevitably affect the disease infection in the near future, which inspire us to include spatial information in disease prediction. In this section, comparative experiments were conducted to empirically investigate whether the inclusion of spatial information can indeed improve the prediction capabilities. For experimental settings that do not consider spatial information, only environmental variables and disease severity of the observation spot are included. The settings of spatio-temporal predictive experiments that consider spatial information are introduced in Section 4. Table 3 demonstrates the results of this comparative experiment group. Prediction performances under different evaluation metrics are listed.

By comparing the performances of prediction models in the upper half and the lower half of Table 3, we can see that with the inclusion of spatial information, all seven models attain better MAE, MAPE and RMSE values than those using only local environmental variables, outperforming them by 0.120 on MAE, 4.18% on MAPE and 0.0195 on RMSE on average. The results of these comparative experiments indicate that the enhanced performances are attributable to the inclusion of complementary spatial information. The overall MAPE values of spatio-temporal prediction models range from 20.24% to 55.20%, with a minimum error attained by the STRNN-EL model. It also has a lowest MAE value of 0.0402 and a lowest RMSE value of 0.0526 which suggests that the proposed method can predict future disease severity at a relatively accurate level. In addition, STRNN-EL model performs better than STRNN model under different evaluation metrics and feature sets. It is convinced that ensemble learning techniques do help improve the generalization ability of the model by keeping more disturbances in the ensemble system. Overall, the results of empirical analysis illustrate the efficacy of STRNN-EL as a viable method for crop disease severity prediction.

5.5 Discussion

The empirical experiments have indicated that the proposed STRNN-based ensemble learning framework for crop disease severity prediction is effective. RNN-based deep learning models show their advantages on time-series modeling. Based on reported cases of wheat yellow rust

Table 3 Performances of various methods using all environmental variables

Models	without spatial information		
	MAE	MAPE	RMSE
MLR	0.1323	59.20%	0.1839
LR	0.1189	45.33%	0.1774
ANN	0.1025	42.32%	0.1702
SVM	0.0971	41.27%	0.1588
RF	0.0906	36.08%	0.1394
STRNN	0.0593	28.54%	0.0743
STRNN-EL	0.0498	26.13%	0.0740
Models	with spatial information		
	MAE	MAPE	RMSE
MLR	0.1262	55.20%	0.1805
LR	0.0968	39.82%	0.1535
ANN	0.0919	39.44%	0.1554
SVM	0.0913	38.49%	0.1412
RF	0.0724	31.24%	0.0925
STRNN	0.0479	25.16%	0.0661
STRNN-EL	0.0402	20.24%	0.0526

المنارة للاستشارات

outbreaks in the Longnan city during 2010 to 2016, we built a specific dataset with relevant environmental variables and disease severity. Comparative experiment groups conducted on this dataset have revealed that the proposed ensemble learning framework facilitates the improvements of disease prediction performances. More concretely, results of comparative experimental group 1 indicate that STRNN model outperformances traditional statistical methods and general machine learning techniques on non-spatial time-series modeling. Comparative experimental group 2 reveals that the inclusion of spatial information can significantly improve the prediction capabilities. Ensemble learning techniques are also proved to be an effective mechanism to improve prediction accuracy and enhance procedural bias.

6 Conclusions and future work

In this paper, we propose a spatio-temporal recurrent neural network model which is an extension of RNN in time and space to predict wheat yellow rust severity. To the best of our knowledge, it is the first time that crop disease prediction tasks are analyzed by a deep learning method. We analyze the patterns of environmental time-series variables by taking advantage of the properties of RNNs. Based on a full understanding of possible risk factors of wheat yellow rust, we include a set of bioclimatic, topographic and soil variables to comprehensively investigate potential relationships between environment conditions and disease severity. Five key remote sensing indexes, including MODIS NDVI, LST, LAI, GPP and NPP, are used as the main time-series features. Bioclimatic, topographic and soil data are extracted from WorldClim, SRTM DEM and ISRIC-WISE database, respectively. Empirical experiments are conducted on a specific dataset which is built based on reported cases of wheat yellow rust outbreaks in the Longnan city during 2010 to 2016. The empirical experiments validate the appropriateness and superior performance of our proposed model on wheat yellow rust severity prediction in China.

There are still some potential problems to be settled in future work. For example, how to retrain the deep learning model in a relatively short time. Applying adaptive learning methods to fine-tune the weights of the deep networks may be a feasible method. Moreover, better prediction accuracy is still being pursued by applying advanced pattern recognition techniques. Determine how to integrate more data sources, such as disease severity analysis results of remote sensing images, into the prediction framework is also a meaningful work. It is believed that the performance of the crop disease prediction system for agricultural emergency management can be improved to a new step with the inclusion of more comprehensive information.

Acknowledgements This work was supported in part by National Natural Science Foundation of China (Grant No. 71301163, 71771212), Humanities and Social Sciences Foundation of the Ministry of Education (No. 14YJA630075, 15YJA630068), Hebei Social Science Fund (HB13GL021), and the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China (No. 15XNLQ08).

References

1. Aydin B, Akkineni V, Angryk R (2016) Mining spatiotemporal co-occurrence patterns in non-relational databases. GeoInformatica 20(4):801–828. https://doi.org/10.1007/s10707-016-0255-0
2. Babaei S, Geranmayeh A, Seyyedsalehi SA (2012) Towards designing modular recurrent neural networks in learning protein secondary structures. Expert Syst Appl 39(6):6263–6274. https://doi.org/10.1016/j.eswa.2011.12.059

المنارة للاستشارات

3. Beed FD (2014) Managing the biological environment to promote and sustain crop productivity and quality. Food Security 6(2):169–186. https://doi.org/10.1007/s12571-014-0333-9
4. Bradley PE, Paul N (2014) Comparing G-maps with other topological data structures. GeoInformatica 18(3):595–620. https://doi.org/10.1007/s10707-013-0191-1
5. Cang S, Yu H (2012) Mutual information based input feature selection for classification problems. Decis Support Syst 54(1):691–698. https://doi.org/10.1016/j.dss.2012.08.014
6. Chakraborty S, Ghosh R, Ghosh M, Fernandes CD, Charchar MJ, Kelemu S (2004) Weather-based prediction of anthracnose severity using artificial neural network models. Plant Pathol 53(4):375–386. https://doi.org/10.1111/j.1365-3059.2004.01044.x
7. Chen G, Wang H, Ma Z (2005) Forecasting wheat stripe rust by discrimination analysis. Plant Prot 32(4): 24–27
8. Chen W, Wellings C, Chen X, Kang Z, Liu T (2014) Wheat stripe (yellow) rust caused by Puccinia striiformis f. Sp. tritici. Mol Plant Pathol 15(5):433–446. https://doi.org/10.1111/mpp.12116
9. Chen X, Liu X, Wang Y, Gales MJ, Woodland PC (2016) Efficient training and evaluation of recurrent neural network language models for automatic speech recognition. IEEE Trans Audio Speech Lang Process 24(11):2146–2157. https://doi.org/10.1109/TASLP.2016.2598304
10. Chherawala Y, Roy PP, Cheriet M (2016) Feature set evaluation for offline handwriting recognition systems: application to the recurrent neural network model. IEEE Trans Cybernetics 46(12):2825–2836. https://doi.org/10.1109/TCYB.2015.2490165
11. Coakley SM, Line RF, McDaniel LR (1988) Predicting stripe rust severity on winter wheat using an improved method for analyzing meteorological and rust data. Phytopathology 78(5):543–550. https://doi.org/10.1094/Phyto-78-543
12. De Wolf ED, Franel LJ (1997) Neural networks that distinguish infection periods of wheat tan spot in an outdoor environment. Phytopathology 87(1):83–87. https://doi.org/10.1094/PHYTO.1997.87.1.83
13. Dos Santos RF, Boedihardjo A, Shah S, Chen F, CT L, Ramakrishnan N (2016) The big data of violent events: algorithms for association analysis using spatio-temporal storytelling. GeoInformatica 20(4):879–921. https://doi.org/10.1007/s10707-016-0247-0
14. El Jarroudi M, Kouadio L, Bock CH, El Jarroudi M, Junk J, Pasquali M, Maraite H, Delfosse P (2017) A threshold-based weather model for predicting stripe rust infection in winter wheat. Plant Dis 101(5):693–703. https://doi.org/10.1094/PDIS-12-16-1766-RE
15. Farquad MAH, Bose I (2012) Preprocessing unbalanced data using support vector machine. Decis Support Syst 53(1):226–233. https://doi.org/10.1016/j.dss.2012.01.016
16. Grabow BS, Shah DA, DeWolf ED (2016) Environmental conditions associated with stripe rust in Kansas winter wheat. Plant Dis 100(11):2306–2312. https://doi.org/10.1094/PDIS-11-15-1321-RE
17. Haghighat M, Abdel-Mottaleb M, Alhalabi W (2016) Discriminant correlation analysis: real-time feature level fusion for multimodal biometric recognition. IEEE Trans Inf Forensics Secur 11(9):1984–1996. https://doi.org/10.1109/TIFS.2016.2569061
18. Hammer B (2000) On the approximation capability of recurrent neural networks. Neurocomputing 31(1): 107–123. https://doi.org/10.1016/S0925-2312(99)00174-5
19. Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. Int J Climatol 25(15):1965–1978. https://doi.org/10.1002/joc.1276
20. Hua T, Chen F, Zhao L, CT L, Ramakrishnan N (2016) Automatic targeted-domain spatiotemporal event detection in twitter. GeoInformatica 20(4):765–795. https://doi.org/10.1007/s10707-016-0263-0
21. Jalal ME, Hosseini M, Karlsson S (2016) Forecasting incoming call volumes in call centers with recurrent neural networks. J Bus Res 69(11):4811–4814. https://doi.org/10.1016/j.jbusres.2016.04.035
22. Kaundal R, Kapoor AS, Raghava GP (2006) Machine learning techniques in disease forecasting: a case study on rice blast prediction. BMC Bioinf 7(1):485. https://doi.org/10.1186/1471-2105-7-485
23. Kumar PV (2014) Development of weather-based prediction models for leaf rust in wheat in the indo-Gangetic plains of India. Eur J Plant Pathol 140(3):429–440. https://doi.org/10.1007/s10658-014-0478-6
24. Landschoot S, Waegeman W, Audenaert K, Van Damme P, Vandepitte J, De Baets B, Haesaert G (2013) A field-specific web tool for the prediction of fusarium head blight and deoxynivalenol content in Belgium. Comput Electron Agric 93:140–148. https://doi.org/10.1016/j.compag.2013.02.011
25. Landschoot S, Waegeman W, Audenaert K, Haesaert G, Baets B (2013) Ordinal regression models for predicting deoxynivalenol in winter wheat. Plant Pathol 62(6):1319–1329. https://doi.org/10.1111/ppa.12041
26. Lawrence S, Giles CL, Fong S (2000) Natural language grammatical inference with recurrent neural networks. IEEE Trans Knowl Data Eng 12(1):126–140. https://doi.org/10.1109/69.842255
27. Luo J, Zhang J, Huang W, Xu X, Jin N (2010) Preliminary study on the relationship between land surface temperature and occurrence of yellow rust in winter wheat. Disaster. Advances 3(4):288–292

المنارة للاستشارات

28. Medeiros CB, Joliveau M, Jomier G, De Vuyst F (2010) Managing sensor traffic data and forecasting unusual behaviour propagation. GeoInformatica 14(3):279–305. https://doi.org/10.1007/s10707-010-0102-7
29. Mehra LK, Cowger C, Gross K, Ojiambo PS (2016) Predicting pre-planting risk of Stagonospora nodorum blotch in winter wheat using machine learning models. Front Plant Sci 7. https://doi.org/10.3389/fpls.2016.00390
30. Oerke EC, Dehne HW, Schönbeck F, Weber A (2012) Crop production and crop protection: estimated losses in major food and cash crops. Elsevier, Amsterdam
31. Paul PA, Munkvold GP (2005) Regression and artificial neural network modeling for the prediction of gray leaf spot of maize. Phytopathology 95(4):388–396. https://doi.org/10.1094/PHYTO-95-0388
32. Pérez-Ariza CB, Nicholson AE, Flores MJ (2012) Prediction of coffee rust disease using bayesian networks. In: Proceedings of the Sixth European Workshop on Probabilistic Graphical Models. DECSAI University of Granada, Spain, pp 259–266
33. Rather AM, Agarwal A, Sastry VN (2015) Recurrent neural network and a hybrid model for prediction of stock returns. Expert Syst Appl 42(6):3234–3241. https://doi.org/10.1016/j.eswa.2014.12.003
34. Sapoukhina N, Paillard S, Dedryver F, Vallavieille-Pope C (2013) Quantitative plant resistance in cultivar mixtures: wheat yellow rust as a modeling case study. New Phytol 200(3):888–897. https://doi.org/10.1111/nph.12413
35. Savary S, Nelson A, Willocquet L, Pangga I, Aunario J (2012) Modeling and mapping potential epidemics of rice diseases globally. Crop Prot 34:6–17. https://doi.org/10.1016/j.cropro.2011.11.009
36. Shi MW (2011) Based on time series and RBF network plant disease forecasting. Procedia Engineering 15:2384–2387. https://doi.org/10.1016/j.proeng.2011.08.447
37. Te Beest DE, Paveley ND, Shaw MW, Van Den Bosch F (2008) Disease–weather relationships for powdery mildew and yellow rust on winter wheat. Phytopathology 98(5):609–617. https://doi.org/10.1094/PHYTO-98-5-0609
38. Tian Y, Klasky S, Abbasi H, Lofstead J, Grout R, Podhorszki N, Liu Q, Wang Y, Yu W (2011) EDO: improving read performance for scientific applications through elastic data organization. In: IEEE International Conference on Cluster Computing. IEEE, Austin, pp 93–102
39. Übeylï M, Übeylï ED (2010) Using recurrent neural networks for estimation of minor actinides' transmutation in a high power density fusion reactor. Expert Syst Appl 37(4):2742–2746. https://doi.org/10.1016/j.eswa.2009.08.005
40. Wakie TT, Kumar S, Senay GB, Takele A, Lenho A (2016) Spatial prediction of wheat septoria leaf blotch (Septoria tritici) disease severity in Central Ethiopia. Eco Inform 36:15–30. https://doi.org/10.1016/j.ecoinf.2016.09.003
41. Wang J, Wang J (2016) Forecasting energy market indices with recurrent neural networks: case study of crude oil price fluctuations. Energy 102:365–374. https://doi.org/10.1016/j.energy.2016.02.098
42. Yue H, Rilett LR, Revesz PZ (2016) Spatio-temporal traffic video data archiving and retrieval system. GeoInformatica 20(1):59–94. https://doi.org/10.1007/s10707-015-0231-0
43. Zhang J, Pu R, Loraamm RW, Yang G, Wang J (2014) Comparison between wavelet spectral features and conventional spectral features in detecting yellow rust for winter wheat. Comput Electron Agric 100:79–87. https://doi.org/10.1016/j.compag.2013.11.001

Dr. Xu is an associate professor at School of Information, Renmin University of China. He is a research fellow at Department of Information Systems, City University of Hong Kong. He got his bachelor and master degree in

المنارة للاستشارات

Mathematics at Xi'an Jiaotong University and doctor degree in Management Science at Chinese Academy of Sciences. His research interests include big data analytics, business intelligence and decision support systems. He has published over 70 research papers in international journals and conferences, such as Decision Support Systems, European Journal of Operational Research, IEEE Trans. Systems, Man and Cybernetics, International Journal of Production Economics, and Production and Operations Management.

Mr. Wang is a Ph. D. student at School of Information, Renmin University of China. He got his bachelor degree in Computer Sciences at School of Information, Renmin University of China. His research interests include smart city, risk management, big data analytics, and decision support systems.

Mr. Chen is a Ph. D. student at School of Information, Renmin University of China. He got his bachelor degree in Telecommunications Engineering with Management at International School, Beijing University of Posts and Telecommunications. His research interests include big data analytics, business intelligence and decision support systems. He has published several papers in international journals and conferences, such as Electronic Commerce Research.

المنارة للاستشارات

GeoInformatica is a copyright of Springer, 2018. All Rights Reserved.

www.manaraa.com