EasyChair Preprint

№ 1197

Building the Rice Blast Disease Prediction Model based on Machine Learning and Neural Networks

Jia-You Hsieh, Wei Huang, Hsin-Tieh Yang, Chia-Chieh Lin, Yo-Chung Fan and Huan Chen

June 15, 2019

Building the Rice Blast Disease Prediction Model based on Machine Learning and Neural Networks

Jia-You Hsieh, Wei Huang, Hsin-Tieh Yang, Chia-Chieh Lin, Yo-Chung Fan, Huan Chen

Department of Computer Science and Engineering, National Chung Hsing University
Taichung city, Taiwan

chayou@gmail.com , U0233147@smail.nchu.edu.tw, 03051021@me.mcu.edu.tw, w103056020@mail.nchu.edu.tw, yfan@nchu.edu.tw, huanchen1107@gmail.com

***Abstract* – Rice blast disease (RBD) is one of the most damaging crop disease for the rice in Taiwan. RBD may be widespread and cause severe losses if it is not controlled in the early stage. The goal of this research is to build an early warning mechanism for the RBD using the machine learning model under current climatic condition. Five years of climatic data (ranging from 2014 to 2018) are used as candidate features in our model, which are collected by the Taiwan government. The RBD conditions are labeled via the field observation during these years. With the climate data, we conduct the recursive feature elimination algorithm to select the key features that have impacts on the RBD. To derive the RBD prediction model, we applied the Auto-Sklearn and neural network algorithms to train the classification model. The experiment results show that the proposed model can classify the RBD conditions (whether exacerbated or relieved) with an accuracy of 72% in average. In particular, our model can achieve an accuracy of 89% in the exacerbation case, which demonstrates the effectiveness of the proposed classification model.**

Keywords- Rice Blast Disease, Recursive Feature Elimination, Neural Network, Auto-Sklearn

I. INTRODUCTION

A. Background information

Rice blast disease (RBD) is a serious crop epidemic, which may be widespread and thus cause huge crop losses. In literature, many experts and the data scientists in agriculture have tried to predict the burst of RBD using various climatic and weather measurements. RBD caused by the fungus Magnaporthe oryzae (formerly Magnaporthe grisea) may need only a few hours to infect the rice. Once infected, the mycelium of the Magnaporthe oryzae will grow and spread rapidly and infect nearby crops. In particular, the rice blast infection cycle is very short. As such, when the climate and environmental conditions are suitable for growth of the fungus, RBD will spread massively in only two or three days. The mycelium of the M. oryzae often attach on the surface of rice leaves and the rice ears. It will damage the crops and reduce the area of photosynthesis. Serious infection will cause crops dead and thus cause economic losses in agriculture.

Fig. 1 shows a typical rice leaf infected by the RBD, which is referenced from the Plant Quarantine Illustration of the Council of Agriculture, Executive Yuan, R.O.C. (Taiwan). The common causes of the RBD include (1) *temperature*, (2) *humidity* and (3) *the duration of sunshine*. They of elaborated in detail as follows.

1) *Cause of temperature:* If the temperature varies a lot in the environment, the resistance ability to the RBD for rice will decrease. As a result, the chance to be infected by fungus will be greatly increased.

2) *Cause of humidity:* In a high-humidity environment, the water film attached on rice leaves will help fungus to grow and spread. Therefore, if the rice leaf stays in high humidity environment for a long time, it will increase the chance to get infected by RBD.

3) *Cause of sunshine duration:* The duration of sunshine has impact on the health of crops. In this study, we will be explore the correlation between humidity and the duration of sunshine.

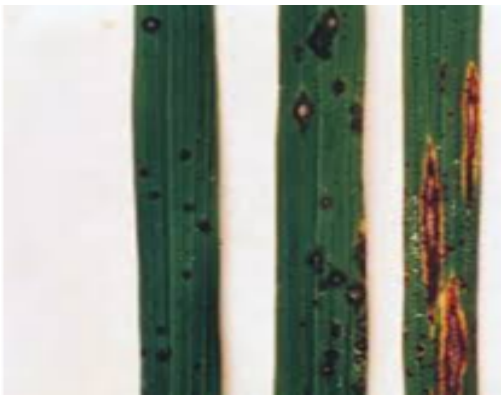

Fig. 1 The typical sample of rice leaf infected by the RBD. This picture is referenced from the Plant Quarantine Illustration of the Council of Agriculture, Executive Yuan, R.O.C. (Taiwan).

.

B. Research purpose

This study is used the statistics of RBD data set provided by the Council of Agriculture, Executive Yuan, R.O.C. We first select many potential features that may cause of the occurrence for RBD. Next, the most important set of key features are selected using the recursive feature elimination methods. We then build our prediction model based on these selected features. To make our model perform best and robust, several classification algorithms are tested. The purpose of the study is to understand which climatic factors will be responsible for the aggravation of RBD. In addition, climate factors derived from the government may combine with the data collected from local sensors in the farmland. As such, it can immediately notify farmers whether the RBD is getting controlled.

C. Literature review

Dated back to 1923, (McKinney, 1923) proposed the Percent Disease Index to predict the occurrence of RBD. The authors used the temperature, humidity and rainfall to establish a multiple regression model to predict RBD. (A.R. Malicdem, 2015) tried to predict the occurrence of rice blast for each rice growing season based on the historical climate of the northern Philippines. The authors used the information of the highest temperature, the lowest temperature, the highest relative humidity, the lowest relative humidity, the average rainfall, the average sunshine, and the average wind speed during the six growth periods of rice as training data. He also applied the neural network and SVM to train the model separately for predicting the rice growth. The results of the study indicated that rainfall is the most important factor that affects the occurrence of rice blast, while the effects on sunshine and wind speed on rice blast are relatively small. In 2017, (Rini et al., 2017) performed this work in western (Odisha & Y. Kim, 2017) performed this work based on the historical climate of South Korea. Base on the data on the incidence of rice blast in the whole year, the RNN training model was used to predict the incidence of rice blast in the next year. The parameters used include the average temperature, average relative humidity and average sunshine hours of the rice growing season.

In literature, many scholars have proposed methods to improve the accuracy of the original prediction by fine-tuning the parameters of the neural network and classification models. (M. Ferrer et al., 2015) used Auto-Sklearn to automatically select a best fit data set classifier from multiple classification models, by automatic fine-tuning the optimization hyper-parameters to improve the final prediction accuracy. The authors of (I Guyon&A. Elisseeff, 2003) focus on whether features selection can improve the performance. Results showed that the filtered features set of importance can effectively improve the prediction accuracy of the machine learning model.

Several machine learning studies have been conducted on agricultural application cases. In (S. Chakraborty et al., 2004), the authors build a model based on neural networks that predict the impact of climate on anthracnose severity. Authors of (K. Klem et al, 2007) analyze climatic factors and deoxynivalenol content in wheat grain based on weather data and preceding crops.

II. METHOD

Rice leaves disease is one of the common rice blast, and may occur in all growth stages of rice. Compared with other RBD, leaf rice blast has a more complete data set, and the classifier established can be applied to all rice growing season. Based on the above factors, we decided to study the correlation of the occurrence of leaf rice blast with the climate and weather related factors. In this section, we will first describe the dataset we used in this study. Then the data processing flow along with each steps will be described.

A. Dataset Description

The data used in this experiment is from the weather-related data and RBD dataset provided by the Agriculture Council, Taiwan (available in public). The data from different zones in Taiwan were collected since 2014 to 2018. To our experience, the dataset used to train the model should be collected from places with similar latitude, since they may have similar weather and rice growth conditions. The experimental data are primarily taken from Yunlin and Chiayi area (They locate in similar latitudes with similar climate conditions). The climate data we collected are categorized into two types: (1) atmospheric climate data are shown in TABLE I (Contains hourly humidity data) and (2) micro-climate data are shown in TABLE II. The atmospheric climate data are obtained from the weather stations and the micro climate data are collected from the temperature and humidity sensors we installed in the farmland. To get the RBD infection labels, research assistants will patrol the farmland weekly and evaluate the RBD condition by hand. For each farmland, the RBD incident rates are also computed and recorded.

TABLE I
ATMOSPHERIC CLIMATE RAW DATA

Date	Max. Temp.	Min. Temp.	Temp. Diff.	Max. Hum.	Min. Hum.	Hum. Diff.
20180409	35.47	12.00	23.47	100	58.48	41.52
20180410	38.19	12.31	25.88	100	58.88	41.11
20180411	38.00	17.06	20.93	100	67.60	32.39
20180412	34.86	20.60	14.26	100	70.60	29.39
20180413	38.58	12.31	17.27	100	71.92	28.07
20180414	37.86	23.66	14.20	100	69.58	30.41
20180415	33.13	17.39	15.74	100	71.87	28.12
20180416	32.97	17.08	15.89	100	80.14	19.85
20180417	30.97	16.48	14.48	100	76.92	23.07

TABLE II
MICRO CLIMATE RAW DATA

Date	Max. Temp.	Min. Temp.	Temp. Diff.	Max. Hum.	Min. Hum.	Hum. Diff.	Rel. Hum.
20180409	27.9	13.2	14.70	99.8	50.6	49.20	76.98
20180410	31.5	14.1	17.40	95.8	47.9	47.90	73.90
20180411	32	19.2	12.80	100	67.3	32.70	85.67
20180412	32.3	22.6	9.70	100	71.3	28.70	86.62
20180413	33.2	22.1	11.10	100	71.1	28.90	88.04
20180414	33.7	25	8.70	100	67.7	32.30	87.25
20180415	28.3	17.8	10.50	100	73.9	26.10	89.98
20180416	26.8	17.6	9.20	100	72.2	27.80	87.23
20180417	25.7	17.6	8.10	99.9	78.9	21.00	92.03

B. Data Processing Flow

This section discusses the data processing flow as shown in Fig. 2. The data processing is performed with the following five steps including (1) Data preprocessing, (2) Label definition, (3) Feature set selection and characteristics analysis, (4) Feature vectors and threshold adjustment, and (5) Classification model selection. These steps are elaborated as below.

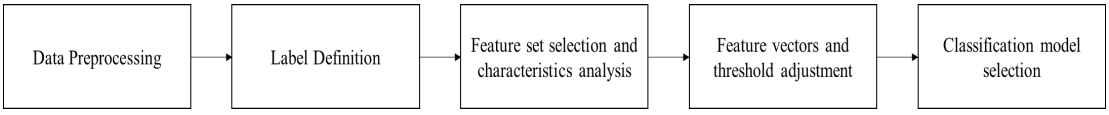

Fig. 2 RBD data processing flowchart

(1) ***Data preprocessing*****:** The first step is to collect the target datasets provided by the Agricultural Commission, Taiwan. Most data are from the Yunlin Chiayi counties, the biggest agricultural production area, in Taiwan. Then we performed the preprocessing on these datasets. The aim is to integrate multiple files into a unified dataset. We also removed those data with incomplete information. In addition to the features on the collected dataset, we also derived several features that

may help in our classification task. Since a big temperature difference will make the rice less resistant to the RBD infection, so we added four features in the datasets, including:

- Feature 1: The temperature difference between today's highest and yesterday's lowest (one day ago).
- Feature 2: The temperature difference between today's highest and the lowest record two days ago.
- Feature 2: The temperature changing rate in the same day, which is defined as in Equation (1). Here, T_h represents the highest temperature, and T_l represents the lowest temperature.
- Feature 4: Humidity infection index, f(H). The humidity infection index is a function of humidity. Experience shows that, high temperature and high humidity may help spores to invade and grow. In particular, if the spores stay in rice for more than 3 hours, the probability of germination will raise up to more than 80%. Based on the aforementioned domain knowledge, we define the feature 4 as in Equation (2).

$$T_r = (T_h - T_l)/T_h \times 100\% \tag{1}$$

$$f(H) = \begin{cases} 1, & H > 80 \ for \ 3 \ consecutive \ hours \\ 0, & otherwise \end{cases} \tag{2}$$

(2) ***Label definition*****:** When researchers patrol and inspect the farmland every week, they will record the current occurrence rate of the RBD every time. They will also compare the newly recorded data with historical data. If the current occurrence rate is higher than the previous one, they will mark "1" indicating that the RBD is spread; otherwise, the marks of "0" will be recorded indicating that the RBD is relieved.

(3) ***Feature set selection and characteristics analysis*****:** According to the records of inspection, we have totally 17 features. (6 atmospheric features (see TABLE I), 7 micro-climate features (see TABLE I), and 4 additional derived features. The labels we use is the trend of RBD infection, the marks made by the agriculture assistants. In our study, we use a recursive feature elimination algorithm to select the most effective features from these 17 features. The selection algorithm first derived the significance of each feature based on the associated weighting returned by the feature selection algorithm. Then we exclude the least important feature at a time, and repeat the above process to determine the final set of featured. In other words, we can compute the importance of each feature to the classifier by evaluating the performance of the algorithm.

(4) ***Feature vectors and threshold adjustment*****:** Some features need to adjust threshold to perform better, which will be discussed here. First of all, we wonder whether the observation period will affect the prediction result. We assume that the number of observation period is set to N. Since the recorder usually inspects once a week, so we will test on the observation period from 1 to 6, to see which parameter performs better. In addition, as the data are recorded based on personal profession of the individual researcher, his/her opinion to the incidence of RBD may not be consistent for every inspection. Since the records is recorded manually, it is assumed that the way of manual recording may cause the incidence of error statistics. We propose a method to adjust the threshold to eliminate errors, and verify whether this method can improve the accuracy of the classifier in predicting RBD. To avoid the error, we used a threshold to remove the outlier. In other words, we will ignore the training data if the difference (in %) between the two incidence rates is less than the threshold, which difference (in %) is calculated as in Equation (3).

$$D = \frac{|D_c - D_p|}{\max(D_c, D_p)} \times 100\% \tag{3}$$

Where D_c is the current incidence of RBD, and D_p is the previous incidence of RBD.

(5) ***Proposed classification models*****:** The aim of our study is to understand which climatic factors will strengthen the occurrence prediction of the RBD. We use neural networks as show in Fig. 3 and Auto-Sklearn to establish the classification model. Machine learning has a wide variety of classification model algorithms, each algorithm has its own characteristics. In order to simplify the process, Auto-Sklearn was applied. In total 15 classification algorithms in TABLE III. In addition, 14 feature pre-processing methods can be selected automatically by the Auto-Sklearn to perform the most suitable pre-processing takes for the dataset and tested classification algorithms, including data regularization and optimization of training parameters. We take the potential features as inputs for the two schemes (Auto-Sklearn and neural network) to forecast the occurrence of RBD.

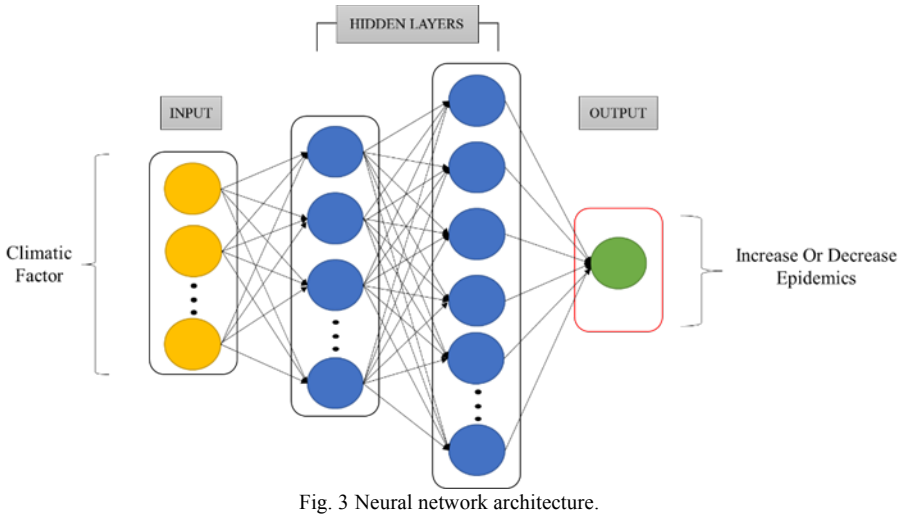

Fig. 3 Neural network architecture.

TABLE III
AUTO-SKLEARN CLASSIFICATION ALGORITHM

Algorithms		
AdaBoost (BA)	LDA	gradient boosting (GB)
Bernoulli naive Bayes	Linear SVM	K Nearest Neighbor
Decision Tree (DT)	Kernel SVM	Linear Class. (SGD)
Extremely Random Trees	Multinomial naive Bayes	QDA
Gaussian naive Bayes	Passive aggressive	Random forest (RF)

III. RESULTS

In this section, we will conduct experiments to show the effectiveness of our proposed scheme. First of all, we compared the effects on RBD for the atmospheric climate features with the microclimate features. The results using are shown in Fig. 4 for the neural network and Auto-Sklearn algorithm respectively.

A. The effect of the relative humidity on the RBD prediction

The results show that the prediction accuracy of RBD using microclimate features is higher than that using the atmospheric climate features under the same conditions. We first compared the performance of the atmospheric climate feature sets. We found that the feature set with the relative humidity information performs better than that without it, which indicates that the relative humidity is a significant factor in RBD prediction. The second finding is that if we combined the microclimate features with the atmospheric climate features in our learning model, the accuracy of the RBD prediction will be increased.

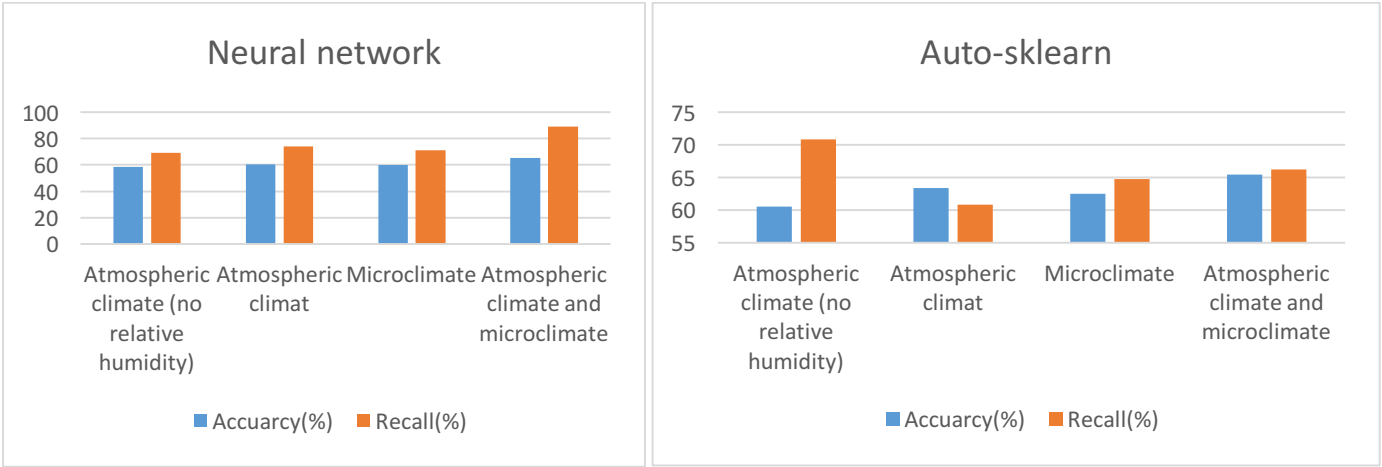

Fig. 4 Neural Network and Auto-Sklearn model index prediction accuracy and recall

B. Effect of the observation duration on the RBD prediction

Next we do experiments on the observation period to see whether it will affect the prediction result. We assume that the number of observation period is set to N, which ranges from 1 to 6. Our goal is to test which parameter performs better. According to the experimental results in TABLE IV, we may find a better setting for N. For the neural network model, it shows that when the value of N is set to 2, the neural network model has 66.6% classification accuracy and a recall of 83.8%. This recall rate is slightly lower than that when N is equals 1, the accuracy rate is the best of all neural network classification models. For the Auto-Sklearn model, if N is equal to 2, the accuracy and recall are 67.4% and 74.5%, respectively. Both metrics are the best of all for the Auto-Sklearn models. Based on the above results, we select feature set A as shown in TABLE V.

TABLE IV
N IN DIFFERENT VALUE MODEL INDEX

	Accuracy (%)		Recall (%)		Precision (%)	
No.	NN	AutoSK	NN	AutoSK	NN	AutoSK
1	62.2	63.9	86.8	68.1	68.7	78.8
2	66.3	67.6	83.8	74.5	82.9	78.9
3	56.7	64.4	54	74.5	78.4	75.2
4	58.6	64.7	64.8	67.8	73.8	78.5
5	62.3	64.5	71.8	68.1	73.9	78.6
6	63.6	63.4	68.1	65.5	77	78.5

TABLE V
FEATURE SET A

Variable	Description	Variable	Description
X_1	Daily maximum temperature in microclimate	X_8	Daily minimum temperature in atmospheric climate
X_2	Daily minimum temperature in microclimate	X_9	Daily temperature difference in atmospheric climate
X_3	Daily temperature difference in microclimate	X_{10}	Daily maximum humidity in atmospheric climate
X_4	Daily maximum humidity in microclimate	X_{11}	Daily minimum humidity in atmospheric climate
X_5	Daily minimum humidity in microclimate	X_{12}	Daily humidity difference in atmospheric climate
X_6	Daily humidity difference in microclimate	X_{13}	Daily maximum relative humidity
X_7	Daily maximum temperature in atmospheric climate	X_{14}	Maximum temperature difference between the minimum temperature of the day and the minimum temperature of the previous day in the microclimate
		X_{15}	Maximum temperature difference between the minimum temperature of the day and the minimum temperature of the previous two days in the microclimate

We used the Recursive Feature Elimination algorithm to sort the significance of features in feature set A. The later the rank (with a bigger value), the less significant for this feature as shown in TABLE VI. We remove the last two less significant features, the X10 and X4 to get the final feature set B.

TABLE VI
RANK OF RECURSIVE FEATURE ELIMINATION

Feature / Year	X_1	X_2	X_3	$\underline{\mathbf{X_4}}$	X_5	X_6	X_7	X_8	X_9	$\underline{\mathbf{X_{10}}}$	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}
2014, 2015	11	1	8	6	5	2	13	9	4	15	14	7	3	12	10
2016	11	4	5	15	6	7	8	12	13	9	14	3	1	10	2
2017	3	1	6	12	8	9	2	7	15	13	11	5	10	4	14
2018	1	2	6	15	8	13	4	3	12	14	5	7	11	9	10
Total	26	8	25	**48**	27	31	27	31	44	**51**	44	22	25	35	36

IV. Conclusion and discussion

In this paper, we tried to solve the RBD prediction problem by using the Machine Learning as well as Neural Network methods. The classification model was built based on the climate and RBD historical data. The accuracy of the predicted results of this classifier is about 72%. This study shows that increasing the number of microclimate factors can increase the prediction accuracy of the RBD. Secondly, the high humidity environment and high temperature are the primary causes of RBD, but the highest humidity of the day for both atmospheric and microclimate does not affect too much for the RBD results. Results also showed that the features from microclimate are more important than those for the atmospheric climate, and the temperature is more important than the humidity.

Acknowledgment

This research was partially supported by the Taiwan Information Security Center at NCHU (TWISC@NCHU), Ministry of Science and Technology, Taiwan, R.O.C., under the grant number MOST 107-2218-E-005-018 and MOST 107-2221-E-005-022.

References

McKinney HH (1923). “Influence of soil temperature and moisture on infection of wheat seedlings by Helminthosporium sativum”. Journal of Agricultural Research, Vol. 26, 195-217.

A. R. Malicdem, & P. L. Fernandez (2015). “Rice blast disease forecasting for northern Philippines”. WSEAS Trans. Inf. Sci. Appl, Vol. 12, 120-129.

P. Rini, M. Dipankar, & B. Naik (2017). “Effect of different meteorological parameters on the development and progression of rice leaf blast disease in western Odisha”. International Journal of Plant Protection, Vol. 10(no. 1), 52-57.

Y. Kim, J.-H. Roh, & H. Y. Kim (2017). “Early forecasting of rice blast disease using long short-term memory recurrent neural networks,” Sustainability, Vol. 10(no. 1), 34.

Chia-Chieh Lin (2018). “Rice Blast Disease Classification based on Weather Factors”. Master’s Thesis of National Chung Hsing University, Taiwan.

M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, & F. Hutter, “Efficient and robust automated machine learning”. Advances in Neural Information Processing Systems, 2962-2970.

I. Guyon & A. Elisseeff (2003). “An introduction to variable and feature selection”. Journal of machine learning research, vol. 3, no. Mar, 1157-1182.

S. Chakraborty, R. Ghosh, M. Ghosh, C. D. Fernandes, M. Charchar, & S. Kelemu (2004). “Weather-based prediction of anthracnose severity using artificial neural network models”. Plant Pathology, vol. 53(no. 4), 375-386.

K. Klem, M. Vanova, J. Hajslova, K. Lancová, & M. Sehnalová (2007). “A neural network model for prediction of deoxynivalenol content in wheat grain based on weather data and preceding crop”. Plant Soil and Environment, vol. 53(no. 10), 421.