Abstract: Rice is a very important food crop of the developing world. However, crop loss is a major concern in rice production, and one of the causes of rice crop loss is the rice blast disease. In this study, we explore how to characterize, in each of the various growth stages of rice crop, the factors that contribute to the occurrence of rice blast disease in selected provinces in the Philippines. Data was gathered from 2 Philippine government agencies, and cleaning and synchronization were performed on the data in order to be able to perform better data analysis. Using the synchronized data, PCA was applied to estimate the relative importance of weather factors in the occurrence of the disease. ANN and SVM models were also built to predict the severity of an occurring rice blast, with the SVM showing a significantly better performance than the ANN. The findings in this study can be used by farmers and other agricultural organizations in order to reduce the occurrence of rice blast disease.

Key–Words: pattern detection, neural network, support vector machine, rice blast, principal component analysis

1 Introduction

Rice is a very important food crop of the developing world. It is the staple food of about half the world’s population. Roughly 900 million of the world’s population depends on rice as producers or as consumers. However, rice crop loss is now one of the major concerns in rice production. Particularly for poor farmers with unfavorable environments, crop loss is a great risk on them and on food security in general [10].

Of the various diseases limiting rice productivity, rice blast disease is an enigmatic problem in several rice growing ecosystems of both tropical and temperate regions of the world and is a serious constraint in realizing the full yield potential of rice cultivars [7]. Rice blast disease causes between 11% and 30% crop loss annually. This represents a loss of about 157 million tons of rice [4].

Considerable efforts have been directed towards developing blast-resistant cultivars. However, due to high variability in blast pathogens, most of the resistant varieties frequently succumb to this disease. The most feasible method for controlling blast epidemics is still the use of fungicides. Unfortunately, due to high cost of chemicals as well as their hazardous effects, use of fungicides is invariably uneconomic. Moreover, farmers are sometimes forced to skip the actual date of fungicide application due to lack of knowledge regarding the actual time of appearance of the disease.

For the judicious use of fungicides, forewarning of blast is important. A major aim of many agricultural forecasting systems is to reduce fungicide use, and accurate prediction is crucial to properly time the application of disease control measures to avoid crop losses [4]. Since weather plays a very important role in the appearance, multiplication and spread of the rice blast fungus, a weather-based forecasting system may provide the desired prediction accuracy [13].

A prediction model based on the relationship between the environmental conditions and the severity of the occurring blast disease could be used to guide management decisions. Thus, if a sound forewarning system is developed, the explosive nature of the disease could be prevented by timely application of the control measures [7]. Such a system will be able to help not only reduce the cost of production but also promote the environmental safety for the operator and consumers by reducing chemical usage.

In this study, we investigate the weather factors that correlate with the occurrence of rice blast disease in selected provinces in the Philippines, a tropical country. We further look at the various growth stages of rice crop and explore the factors that may be used for building predictive models. Finally, we create ANN and SVM models for predicting the severity of an occurring rice blast disease, and statistically compare the performances of these 2 models.
2 Related Literature

2.1 Machine learning techniques in plant disease forecasting

The study of Kaundal et al. stated that diverse modeling approaches such as neural networks and multiple regression have been followed to date for disease prediction in plant populations. There is a need for exploiting new prediction softwares for better understanding of plant-pathogen-environment relationships because of their inability to predict values from unknown data points and longer training times. They also introduced an online page that provides the severity of the plant disease [7].

2.2 Detecting plant health condition

In another study about plant health condition, it was stated that detecting plant health condition is an important step in controlling disease and insect stress in agricultural crops [8]. In discriminating and classifying different fungal infection levels in rice (Oryza sativa L.) panicles, the researchers used artificial neural network and principal components analysis techniques. Four infection levels in rice panicles were used in the study: no infection condition, light and moderate infection caused by rice glume blight disease, and serious infection caused by rice false smut disease. Their results indicated that it is possible to discriminate different fungal infection levels of rice panicles under laboratory conditions using hyperspectral remote sensing data.

2.3 Classifying Diseases of Infected Plants

Phadikar et al. previously developed an automated system for classifying diseases of infected rice plants [12]. They extracted features from the infected regions of the rice plant images by using a system that classifies different types of rice disease. They used Fermi energy based segmentation method to isolate the infected region of the image from its background. Features like colour, shape and position of the infected portion and extracted by developing novel algorithms are the symptoms of the diseases based on the field experts’ opinions. Dubrawski et al. also stated that to reduce complexity of the classifier, important features are selected using rough set theory (RST) to minimize the loss of information [5]. Finally using selected features, a rule base classifier was built that cover all the diseased rice plant images and provides superior result compare to traditional classifiers.

2.4 Collection of data to assess various aspects of rice production

Hirai et al. developed an analytical method that can identify the determinants of rice yield and quality [6]. They assessed the data from 82 paddy fields where rice is produced in various environments and with various management styles by using applied cluster analysis (Ward Method). Initially, the 82 paddy fields were classified into 11 clusters based on five indicators of yield components and rice quality; number of panicles, number of spikelets, percentage of ripened grains, 1000-grain weight (GW) and protein content of brown rice.

In a succeeding study, 9 of the 11 clusters (two clusters were excluded due to insufficient data to form a cluster) were divided into four groups based on yield capacity [3]. They successfully extracted common characteristics from each cluster such as fertilizer application, meteorological environment and growth conditions. Furthermore, determinants of yield components and protein content were efficiently identified based on the common characteristics extracted.

3 Methodology

3.1 Data Acquisition

The data about rice blast occurrence for this research was provided by Department of Agriculture (DA) Regional Unit 1 which is the principal agency of the Philippine government responsible for the promotion of agricultural development. The attributes of each report are the following:

- Report Date - the date of the cropping season when rice blast occurred
- Province - its specific Province
- Municipality/City - its specific Municipality/City
- Barangay - its specific Barangay
- Pest Observed - the kind of Rice Disease occurred
- Rice Crop Variety - the rice varieties approved by the National Seed Industry Council (NSIC), with rice blast resistance level (e.g., NSIC Rc 152 - Intermediate resistance to blast = 3)
- Growth Stage of the Plant - the stage when the rice blast occurred. The possible stages are (1) Seedling, (2) Tillering, (3) Maximum Tillering, (4) Panicle Initiation, (5) Heading/Flowering, (6) Milking and (7) Ripening.
• Severity - the percentage of harshness of the rice blast within the area

Also, history of weather conditions was provided by the different Agromet and Weather stations of Philippine Atmospheric, Geophysical and Astronomical Services Administration (PAGASA) based from the following stations:

• Sinait PAGASA Station
• Dagupan PAGASA Station
• DMMMSU Agromet Station
• MMSU Agromet Station

The attributes of historical weather data gathered from January 2006 to December 2011 are the following:

• Temperature (Max) - maximum temperature averaged for the month (in Celsius)
• Temperature (Min) - minimum temperature averaged for the month (in Celsius)
• Humidity - average amount of vapor in the air for the month (in percentage)
• Rainfall - the average precipitation count for the month (in millimeter)
• Solar Radiation - average radiant energy emitted by the sun (in kW/m²)
• Evaporation - vaporization of a liquid per month (in millimeter)
• Brightness - average brightness per month (in minutes)
• Wind Speed - average air speed per month (in km per day)

3.2 Data Processing

The data acquired from DA and PAGASA were combined into 1 table using synchronization. Specifically, for each row of the DA data describing a specific rice blast report, the corresponding weather information from PAGASA was appended. This resulted in a data set consisting of 122 rows and 16 columns.

After the synchronization, feature scaling was applied to make sure that the features are in similar scale. The mean normalization was applied to scale the features, where the general rule is to take a feature \( x_i \) and replace it with

\[
x_i = \frac{x_i - \mu_1}{s_1}
\]

Where \( \mu_1 \) is the average of the values \( x_i \) of a column in the data set and \( s_i \) is standard deviation of values in \( x_i \). The resulting scaled features values were in the range \(-3 \leq x_i \leq 3\).

Using the resulting synchronized data, Principal Component Analysis (PCA) was used to estimate the most important weather features in the occurrence of rice blast. PCA is a statistical procedure that uses orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.

After we determined the most important features, we characterized the factors that affect the occurrence of rice blast within each Growth Stage of the Plant (GSP). Such characterization may be used in giving advice to farmers and other agricultural operators on when it would be wise to apply some interventions to reduce the risk of rice blast.

3.3 Machine Learning Algorithms for the Prediction of Severity

3.3.1 The Artificial Neural Network

ANNs offer a computational approach that is quite different from conventional digital computation. Mehrotra stated that digital computers operate sequentially and can do arithmetic computation extremely fast [9]. Biological neurons in the human brain are extremely slow devices and are capable of performing a tremendous amount of computation tasks necessary to do everyday complex tasks, commonsense reasoning and dealing with fuzzy situations. The underlying reason is that, unlike a conventional computer, the brain contains a huge number of neurons, information processing elements of the biological nervous system, acting in parallel. ANNs are thus parallel, distributed information processing structures consisting of processing elements interconnected via unidirectional signal channels called connection weights. Although modeled after biological neurons, ANNs are much simplified and bear only superficial resemblance. Some of the major attributes of ANNs are: (1) they can learn from examples and generalize well on unseen data and (2) are able to deal with situations where the input data are erroneous, incomplete or fuzzy [11].

The process of finding the best set of weights for the neural network is referred to as training or learning. The approach used by most software to estimate the weights is backpropagation. Each time the network cycles through the training data, it produces a predicted value for the target variable. This value is compared to the actual value for the target variable, and an error is computed for each observation. The er-
errors are fed back through the network and new weights are computed to reduce the overall error. Despite the neural network terminology, the training process is actually a statistical optimization procedure. Typically, the procedure minimizes the sum of the squared residuals.

From this model the interval activity of the neuron can be shown to be:

$$V_k = \sum_{j=1}^{P} W_{kj} X_j$$

(2)

Figure 1 shows that the output of the neuron, $y_k$, results from some activation function applied to the resulting value of $V_k$, computed using Eq. 2. In this study, the activation function used was a sigmoid function.

### 3.3.2 The Support Vector Machine

Support Vector Machine (SVM) is a family of algorithms that have been implemented in classification, recognition, regression and time series. SVM originated as an implementation of Vapnik’s (1995) Structural Risk Minimization (SRM) principle to develop binary classifications. SVM emerged from research in statistical learning theory on how to regulate generalization, and find an optimal trade-off between structural complexity and empirical risk. SVM classify points by assigning them to one of two disjoint half spaces either in the pattern space or in a higher-dimensional feature space. The main idea of SVM is to construct a hyperplane as the decision surface such that the margin of separation between positive and negative examples is maximized [2].

### 3.4 Data analysis

To generalize the results for each machine learning algorithm, the sample from the population underwent cross validation which is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. The idea behind $k$-fold cross-validation is to divide all the available data items into roughly equal-sized sets. Each set was used exactly once as the test set while the remaining data was used as the training set [1]. In this study, we used $k = 10$.

For performance with respect to predicting severity, the Mean Squared Error ($MSE$) and the Coefficient of Determination which is also known as $R^2$ value was computed.

Mean squared error quantifies the difference between values implied by an estimator and the true values of the quantity being estimated. Its performance is computed using the following formula:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (X_i - Y_i)^2$$

(3)

where $X$ is the vector of $n$ predictions and $Y$ is the vector of true values. Lower MSE values are better, and zero means no error.

Likewise, the coefficient of determination, $R^2$, is simply the square of correlation coefficient, but it is very useful because it gives the proportion of the variance (fluctuation) of one variable that is predictable from the other variable. It is a measure that allows us to determine how certain one can be in making predictions from a specific model. The coefficient of determination is such that $0 \leq R^2 \leq 1$, and denotes the strength of the linear association between $X$ and $Y$. It represents the percent of data that is the closest to the line of best fit. It is computed using Eq. 4

$$R^2 = 1 - \frac{SS_R}{SS_T}$$

(4)

where $SS_R$ is the regression sum of squares and $SS_T$ is the total sum of squares.

### 4 Results and Discussion

This section shows the results and some analysis of the feature selection, quantitative description/characterization of each growth stage and performance outputs of the two machine learning algorithms.
4.1 Feature Selection

From the results of the PCA which was used to explore the most important features that affect the occurrence of rice blast, Rainfall accounts for 46% importance in the occurrence of rice blast, with Temperature Minimum (31%), Temperature Maximum (20%) and Humidity (2%) as the next most important features (see Fig. 2). On the other hand, Solar Radiation, Evaporation, Brightness and Wind Speed have very low and/or minimal influence (<1%), and can therefore be omitted.

With this, the most important weather features aggregated per month are synchronized to the rice disease reports which resulted to a 122 x 5 matrix that describes the synchronized data set. Specifically, the following attributes were used in the characterization of factors for each growth stage and predicting the severity when the rice blast occurred:

\[ x = [x_1, x_2, x_3, x_4, x_5] \] (input matrix)

- \( x_1 \) = Rice Crop Variety Blast Resistance Level
- \( x_2 \) = Rainfall
- \( x_3 \) = Relative Humidity
- \( x_4 \) = Temperature (Max)
- \( x_5 \) = Temperature (Min)

4.2 Growth Stage Risk Characterization

From the synchronized data set from weather and rice blast reports, a simple histogram in Fig. 3 shows that Growth Stage 2 (Tillering) is the most prone stage to rice blast with more than 45% of the total reported occurrences, while Stage 7 (Ripening) has no occurrence of rice blast. With this information, a sound characterization of each growth stage can be explored further. This can help farmers and agricultural organizations to decide on when would it be effective to apply intervention to minimize the possibility of occurrence of rice blast.

Tables 1 and 2 show quantitative descriptions for each stage for the most important factors in the occurrence of rice blast, wherein the maximum and minimum values are revealed. For Stage 1 (Seedling), rice blast will most likely occur if Rainfall is within 14.1-36.4mm, temp. min. is within 23.6-24.1°C, temp. max. is 31.5-33.1°C, and Humidity is 86-87%. Non-occurrence of rice blast observed Rainfall was in minimum count that has an average of 3.56 days per month, while temperature is much higher and Humidity decreased to about an average of 83%.

In Stage 2, which is the most prone in rice blast, the range of Rainfall is within 0-32.4 mm, Temp. Min. is within 16.9-24.9°C, Temp. Max. is 30.8-33.1°C, and Humidity is 83-87%. For the years with non-occurrence of rice blast, an average Rainfall count of 0.35, and Humidity at an average of 83% was observed which are lower than the minimum values.

For Stage 4, Rainfall is within 10.5-32.4mm, Temp. Min. is within 16.9-24.6°C, Temp. Max. is 30.8-32.2°C, and Humidity is 83-86%. In the cropping season with no occurrence of rice blast in this stage, it was observed that Rainfall again was below the minimum count with an average of 5.66 days per
Table 1: Characterization of factors for each Growth Stage of the Plant - Humidity and Temperature Min

<table>
<thead>
<tr>
<th>GSP</th>
<th>Rainfall</th>
<th>Temp Min</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Min</td>
<td>Max</td>
</tr>
<tr>
<td>1</td>
<td>14.1</td>
<td>36.4</td>
</tr>
<tr>
<td>2</td>
<td>5.6</td>
<td>32.4</td>
</tr>
<tr>
<td>3</td>
<td>0.0</td>
<td>29.6</td>
</tr>
<tr>
<td>4</td>
<td>10.5</td>
<td>32.4</td>
</tr>
<tr>
<td>5</td>
<td>0.0</td>
<td>12.1</td>
</tr>
<tr>
<td>6</td>
<td>0.0</td>
<td>19.4</td>
</tr>
</tbody>
</table>

Table 2: Characterization of factors for each Growth Stage of the Plant - Temperature Max and Humidity

<table>
<thead>
<tr>
<th>GSP</th>
<th>Temp Max</th>
<th>Humidity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Min</td>
<td>Max</td>
</tr>
<tr>
<td>1</td>
<td>31.5</td>
<td>33.1</td>
</tr>
<tr>
<td>2</td>
<td>30.8</td>
<td>33.1</td>
</tr>
<tr>
<td>3</td>
<td>30.8</td>
<td>33.1</td>
</tr>
<tr>
<td>4</td>
<td>30.8</td>
<td>32.2</td>
</tr>
<tr>
<td>5</td>
<td>30.8</td>
<td>33.3</td>
</tr>
<tr>
<td>6</td>
<td>31.0</td>
<td>33.2</td>
</tr>
</tbody>
</table>

4.3 Prediction of Severity using ANN and SVM

Table 3 shows the comparison of the ANN and SVM for prediction of rice blast severity using Mean-Squared Error (MSE) and Coefficient of Determination ($R^2$) based on the results of a 10-fold cross-validation. For the MSE, the algorithm with a value closer to zero gains an advantage over the other algorithm. A paired t-test of the cross-validation results indicate that the MSE value for SVM is significantly lower (p-value < 0.001) than that of ANN.

The coefficient of determination gives the proportion of the variance (fluctuation) of one variable that is predictable from the other variable, where an $R^2$ of 1 indicates that the regression line perfectly fits the data. In this measure, the SVM result is likewise better than that of ANN. However, applying a paired t-test shows that the difference is not significant (p-value = 0.105).

Table 3: Mean squared error (MSE) and Coefficient of determination ($R^2$) performance of ANN and SVM

<table>
<thead>
<tr>
<th>Fold Num</th>
<th>ANN</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MSE</td>
<td>$R^2$</td>
</tr>
<tr>
<td>1</td>
<td>0.6207</td>
<td>0.2586</td>
</tr>
<tr>
<td>2</td>
<td>0.6485</td>
<td>0.4853</td>
</tr>
<tr>
<td>3</td>
<td>0.6229</td>
<td>0.6678</td>
</tr>
<tr>
<td>4</td>
<td>0.4614</td>
<td>0.3430</td>
</tr>
<tr>
<td>5</td>
<td>0.7019</td>
<td>0.4474</td>
</tr>
<tr>
<td>6</td>
<td>0.5270</td>
<td>0.6254</td>
</tr>
<tr>
<td>7</td>
<td>0.4815</td>
<td>0.4757</td>
</tr>
<tr>
<td>8</td>
<td>0.4932</td>
<td>0.8521</td>
</tr>
<tr>
<td>9</td>
<td>0.6543</td>
<td>0.0675</td>
</tr>
<tr>
<td>10</td>
<td>0.4995</td>
<td>0.1126</td>
</tr>
<tr>
<td>Average</td>
<td>0.5711</td>
<td>0.4335</td>
</tr>
</tbody>
</table>

5 Conclusions

The most important weather features that affects the occurrence of rice blast are Rainfall, Temperature Minimum and Maximum, and Humidity. Each Growth Stage of the Plant (GSP) has its own characteristic for each feature that determines whether rice blast is likely to occur or not. Also, we found out that Stage 2 (Tillering) is the most prone stage. Based on these observations, the non-occurrence of rice blast has feature values which were lower than the minimum value for the most important weather features. Likewise, the SVM provides a more accurate prediction of the severity of the rice blast with an MSE of 0.3807. These information can help the farmers and other agricultural organizations in determining when it would be wise to have interventions (e.g., application of fungicides) and the associated severity risks for non-intervention. Additional data and other processing techniques can be included for a greater generalization of the prediction models.

References:


[3] Fast Pattern Detection Using Normalized Neural Networks and Cross-correlation in the Fre-


