PADDY LEAF DISEASE PREDICTION USING MACHINE LEARNING TECHNIQUES

Siriki Puspalatha[1], Ch. Satyanand Reddy[2]
[1]PG Scholar Department of Computer Science& Systems Engineering
Andhra university college of engineering (A)Andhra university

Abstract- *Plants are considered the most important resource in the fight against global warming, but they are now threatened by various diseases. Recent research has been done to identify and understand plant diseases. The present paper is intended to detect the diseases in paddy crops, specifically targeting Brown Spot Disease, Leaf Blast Disease, and Leaf Blight Disease - Tungro. These diseases prevent the growth and protection of paddy plants at varying development levels. We suggest here the methodology of a collection of four types of paddy diseases and a cluster of sound paddy leaves. Bacteria, fungi, and other agents cause these paddy diseases. In this respect, a technique was proposed that would be automated to decrease the noise and upgrade the manpower time for the assessment of the impression of paddy leaf diseases. This research has been aimed at achieving the optimal detection of paddy leaf diseases via machine learning techniques and a fully automated detection method. K-Fold cross-validation was employed to measure the classification performance. The paddy leaves were classified under four classes used in models like Random Forest and Support Vector Machine (SVM). Among these, the highest accuracy of 97% was achieved for Random Forest using K-Fold cross-validation to predict the four classes of paddy leaf diseases along with healthy paddy leaves and even create a web application for this project.*

I. INTRODUCTION

Plant diseases are serious environmental and agricultural hazards. Food security and the sustainability of agriculture can be negatively impacted by these diseases, which can significantly impede plant growth and output. Because they absorb carbon dioxide and maintain ecosystem balance, plants are crucial to reducing the effects of climate change. But in recent times, their vulnerability to a range of illnesses has raised concerns, with the agriculture industry being particularly affected by a marked rise in the frequency of plant diseases. This increase has led academics and farmers to look for efficient ways to control and prevent disease.
Farmers face numerous challenges in selecting the best crops and appropriate pesticides to protect their plants. The complexity of modern agriculture requires a deep understanding of plant health and disease dynamics. Plant disorders occur when regular functions are disrupted by soil issues, environmental stresses, or physical factors. These disorders can cause significant damage to crops, leading to reduced yields and economic losses. Unlike plant disorders, diseases caused by viruses, bacteria, and fungi can spread between plants, affecting different parts of the plant above and below the ground. This transmissibility makes managing plant diseases particularly challenging.

The following is our primary contribution to this work:

- Our goal was to diagnose paddy leaf disease using machine learning techniques (RF, SVM) and determine which classifier performs better in disease detection.
- Stratified K-fold Cross-Validation is utilized for classification problems. Each fold in the method is selected to include roughly the same ratios of the target class.
- When new images are given as the input for the system, it predicts the type of disease and helps to take contour action before the plants get affected more.
- The nation's agricultural sector is its backbone, and our efforts will help to maximize field productivity by early plant disease detection.

2. LITERATURE SURVEY AND RELATED WORK

Significant research has been conducted in the field of rice plant disease analysis, focusing on various machine-learning techniques to identify and classify different diseases. This growing body of work underscores the importance of leveraging advanced technologies to address agricultural challenges, particularly those related to plant health and productivity. In India, almost 70% of people work in agriculture. As a result, the company should be fully committed to new work. In India, a diverse range of harvests are being produced. Farmers have been managing their crops using traditional methods, and most of the time, diseases have an impact on the yields, causing a decline in agricultural production. The important argument is that "treating plant diseases with excessive use of pesticides increases costs and increases the possibility of toxic build-up on rural objects." Most importantly, it brings farmers disasters that affect their prosperity.

Archana KS and Sahayad have utilized K-means clustering to detect and segment brown spot and bacterial leaf blight in rice plants. Their work primarily focused on applying segmentation techniques to isolate the affected areas on paddy leaves using K-means clustering. This method allowed for the effective differentiation of healthy and diseased leaf regions, providing a foundation for further analysis and classification.

In 2001, Md. Ashiqul Islam et al. from Daffodil International University published a study on using deep learning with Convolutional Neural Networks (CNNs) for paddy leaf disease detection and classification. This research included four classes: one healthy leaf class and three disease classes. Their process involved image acquisition, image pre-processing, and the application of CNN models for classification, achieving high accuracy in predictions. The use of CNNs marked a significant advancement in the field, given their ability to automatically extract features from images without manual intervention.

Minu Eliz Pothen and Dr. Maya L Pai focused on classifying four paddy disease classes (Bacterial leaf blight, Leaf smut, and Brown spot) using Support Vector Machine (SVM) algorithms. They employed Otsu's method for segmentation and used features extracted with Histogram of Oriented Gradients (HOG) and Local Binary Patterns (LBP). Their research demonstrated that the SVM algorithm with a polynomial kernel function could effectively identify various rice leaf diseases. The combination of HOG and LBP features provides data robust representation of the leaf images, enhancing the accuracy of the SVM classifier.

3. PROPOSED METHODOLOGY

Machine learning is meant to enable computers to learn from data and make decisions or predictions without being explicitly programmed for every specific task. I have proposed a system that classifies paddy leaf diseases as bacteria lb light, blast, brown spot, Tungro, and normal leaves using machine learning and a pre-trained model for the classification task. To do the task of image recognition, CNN is used on a greater scale in the deep learning environment. The flow of work begins with the paddy leaf disease data set and with the process of multiple class classification using machine learning and classifiers to make detection by predicting the given data set.

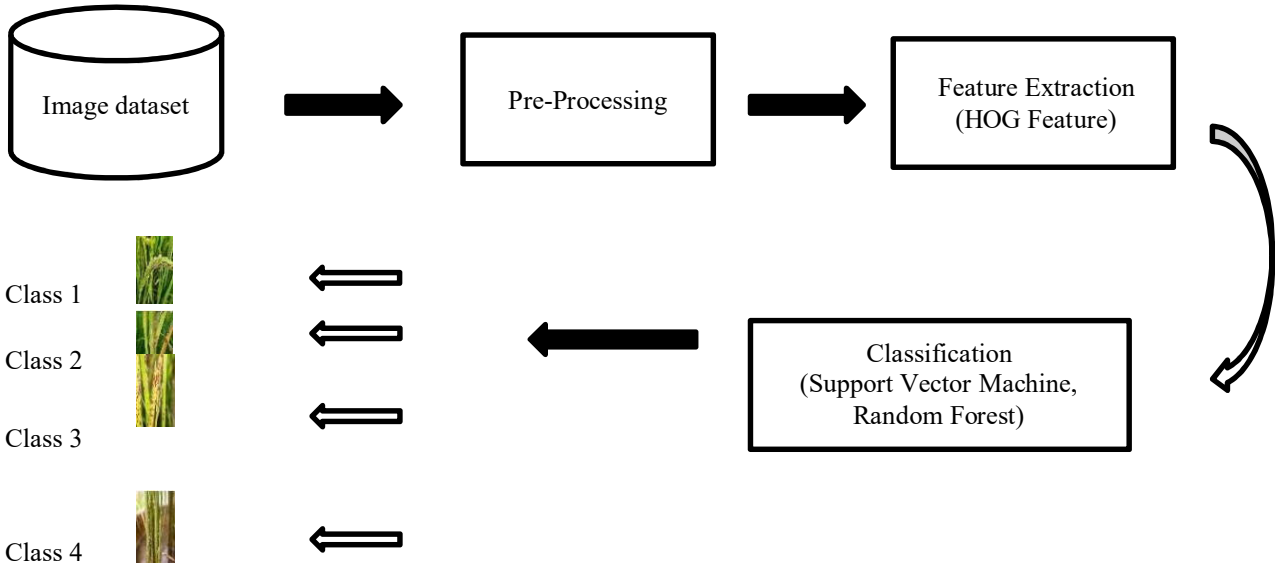


Fig.1.The workflow of Paddy leaf disease prediction.

3.1 Paddy leaf disease dataset:

Images were collected from the field using phones and digital cameras, and some were sourced online. An agricultural officer labeled the entire dataset, resulting in a collection of 5932 images divided into four classes: Brown Spot, Leaf Blast, Leaf Blight, Tungro, and Healthy Leaf, with each class containing 1000 above images.

Table I: Dataset Table

Dataset Table	Image Data Quantity
Total images	5932
Bacterial blight	1584
Blast	1440
Brown spot	1600
tungo	1308

.

Fig. 2. Paddy Leaf Diseases (a) Bacterial blight, (b) Blast,
(c) Brown spot and (d)Tungro

3.2 Image Pre-processing

To prepare the dataset for prediction, several pre-processing steps were undertaken, including scaling, rotating, and flipping images within the same category. This process increased the volume of relevant data and ensured consistency. Images were converted to grayscale, RGB, and HSV formats. Data augmentation techniques were used to enhance the dataset, and images were resized to 224x224 pixels and converted to jpg format. Duplicate and erroneous images were removed.

3.3 Feature Extraction

Feature extraction involves building up image features such as color, shape, texture, and Histogram of Oriented Gradients (HOG) for training and testing. This research focused on global feature extraction to improve training and testing speed by removing extraneous characteristics from the input data. The images were converted to RGB and HSV formats, which play significant roles in achieving better accuracy. The form of the image was described by measuring the width and height in pixels. Four types of global feature extraction were used:

- Colour - Shape
- Texture
- Histogram of Oriented Gradients (HOG)

3.3 Multi-Class classification

Every sample in the dataset will have a single class label assigned to it. The classification procedure assigns a target to the record based on the values of its features. Many algorithms were created from an ML perspective to operate on the training dataset, produce a model, and then predict the class for the testing dataset. Repeated training of the model with varied combinations is the success factor in the data mining arena. Five class labels are classified in our study based on the attributes of the photos. Random sampling techniques and k-fold cross-validation are two of the many ideas that go into selecting the training and testing data. Using the previous approach, k equal-sized subsamples are randomly selected from the original sample. Out of the k subsamples, one subsample is kept as validation data for testing the model, and the remaining k-1 subsets are used as training samples. The subsequent random technique separates the testing and training phases and is based on ratio.

3.3.1 Support Vector Machine (SVM)

A supervised learning model analyzes data for classification and regression analysis. The popular classification algorithm plays an important role in machine learning. Here a simple algorithm classifies the SVM according to given data sets.

Classify a data set into two categories by finding the best separate line or by a hyper plane that which divider the classes.

Consider a data set of n data points each represented by a features vector xi and a corresponding label yi € {-1, 1}, meaning of two classes.
Prepare a dataset D = {(x1y1), (x2y2)….. (xnyn)} Where xi =features vector for the ith data point yi = class label which is either ±1
Define a hyperplane by using wTx+b=0
Where wT is the weight vector which is perpendicular to the hyperplane b is the bias term which shifts the hyperplane away from the origin.
now maximize the distance between the classes i.e., to max. the margin uses the optimization i.e., min ½ ||w||2 subjected to Yi (wTxi+b) ≥1
Now introduce a slack variable εi to optimize the real-world data because it's not always possible to perfectly separate classes.
Optimize the data min ½ || w||2+ +$c\sum i=1n\varepsilon i$
Where C is the regularization parameter which controls the trade-off between maximizing the margin and which allows miss classifications.
To find the optimal values of w and b use sequential minimal optimization (SMO) which separates the hyper-plane.
For a new data point x1 use a prediction
y = sing(wTx+b) i + if y=+1 point belongs to class1 else y= -1 point belongs to class y=-1

3.3.2Random Forest

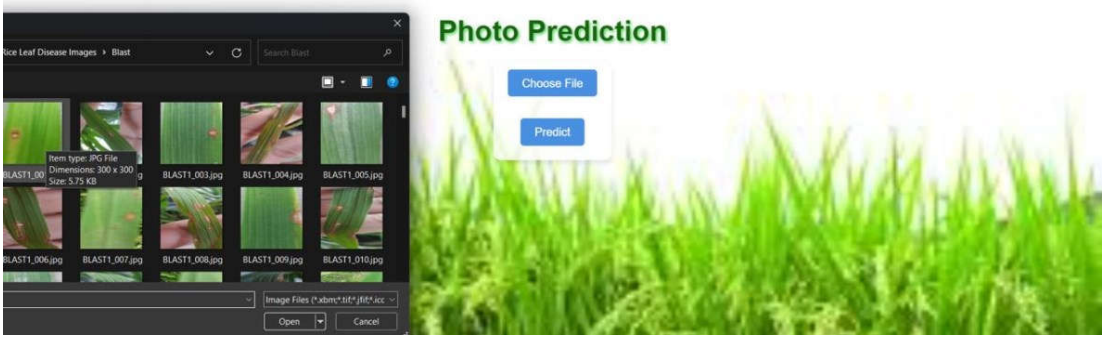
The Random Forest (RF) classifier uses an ensemble learning approach for classification, where the final prediction is based on the averaged outputs of multiple decision trees. During training, several trees are generated to form a "forest," with each tree trained on a random subset of data. Unlike traditional decision tree algorithms that rely on a fixed set of rules and measures like the Gini Index (GI) or information gain, RF classifiers take a different approach by randomly selecting features for the root node's split.
Each tree independently makes a prediction, and the final classification is determined by a majority vote among the trees. This method is widely applied across various fields, including spectral imaging, ecology, and land cover classification. The key parameters that influence the model's performance are the number of trees and growth control settings. In our work, we used 10 trees and ensured that the subset used for splits would not drop below five, ensuring reliable training and reproducibility in the process.

K-Fold Cross-Validation: Using K-fold cross-validation to divide the dataset into K subsets, ensuring each subset is used for both training and testing.
1. Dataset Split: Splitting the dataset into K equally sized folds.
2. Training and Validation: Iteratively using K-1 folds for training and the remaining fold for validation.
3. Performance Metrics Calculation: Averaging the results of each iteration to get an overall performance metric.
Performance Evaluation: Measuring the accuracy, precision, recall, and F1-score of each mode

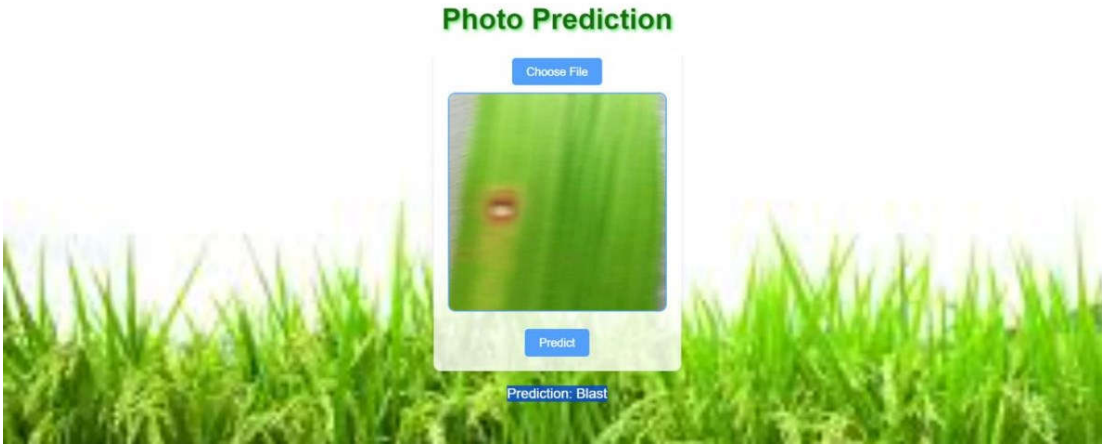
Accuracy: The ratio of correctly predicted instances to the total instances.
Precision: The ratio of correctly predicted positive observations to the total predicted positives. Recall: The ratio of correctly predicted positive observations to all observations in the actual class.
F1-Score: The weighted average of Precision and Recall.
Algorithm Selection: Comparing the performance metrics to select the best-performing model.

A paddy leaf disease web application is a tool designed to identify and diagnose diseases affecting paddy leaves using web technologies.

1. **User Interface**: Paddy leaf photos are uploaded by users using a web interface.



2. **Image Processing and Analysis**: Image analysis techniques, including feature extraction and machine learning models trained to identify different leaf diseases, are used to upload photos.
3. **Disease Classification**: The application uses a pre-trained model to classify the disease based on the extracted features from the leaf images. The model predicts the type of disease and provides feedback to the user.
4. **Results Display**: The application displays the results



4. EXPERIMENTAL RESULT AND ANALYSIS

By employing various ML techniques and K-fold cross-validation, we successfully achieved our objectives. Two algorithms were utilized in this research: Support Vector Classifier (SVC), and Random Forest. Before commencing our work, thorough algorithm selection and dataset labeling by an agriculture officer were performed. After applying K-fold cross-validation techniques across these two models, the highest accuracy was attained with Random Forest, achieving an accuracy score of 97%. This stands as the most significant result for this dataset, showcasing high accuracy.
Across the four classes of paddy leaf diseases.

Table II: Accuracy Table

Classifier	Accuracy Score (%)
Random Forest	97.22
SVM (SVC)	96.80

4.1 Confusion matrix:

The dimension of the confusion matrix is equal to the number of classes. Thus, a 4-class model produces a 4x4 confusion matrix, showing fine details about the mapping of correct and incorrect classifications. It helps in understanding different metrics that help in setting system trends, with every instance being classifiable by the techniques applied. Columns represent actual class labels and rows represent predicted class labels. In other words, every cell inside the matrix is one of these— TP (True Positive), TN (True Negative), FP (False Positive), or FN (False Negative)— based on the comparison of the class it holds against the predicted value. This is where the predicted positive class coincides with the actual positive class, and it classifies the result as TP, or a match between negative predictions and negatives, which is classified as TN. In a case where the model misclassifies by predicting a positive class for a negative instance, then it gets an FP label. And in the opposite sense, where the actual class is positive and the model predicts a negative class, then it's counted as FN. These parameters of TP, TN, FP, and FN will be used for several performance measures quantifying the model's ability

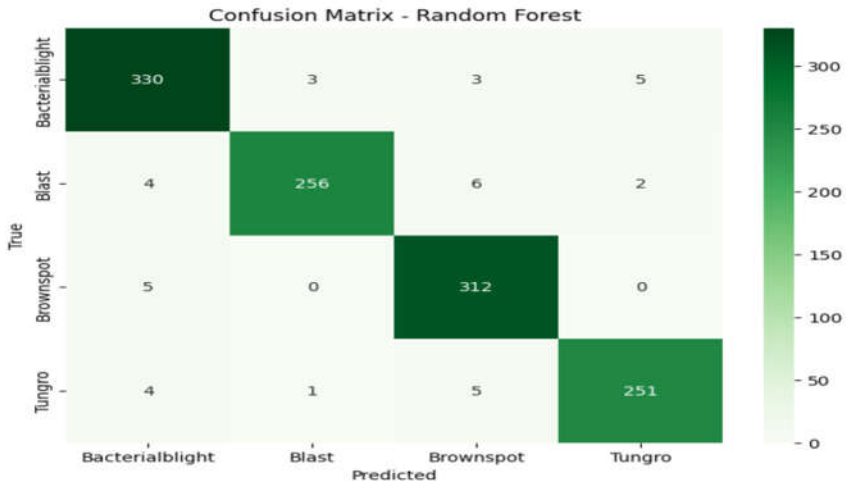


Fig.3. Confusion Matrix of Random Forest

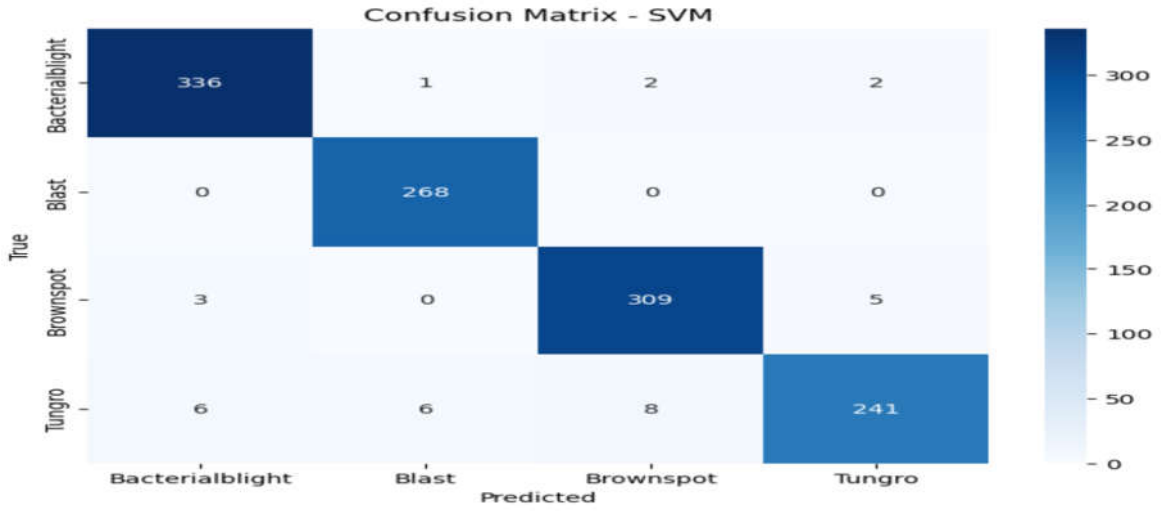


Fig.4. Confusion Matrix of SVC

The figures and matrices depict detailed classification and normalization for achieving the highest accuracy using Random Forest.

4.2 Models Performance Measurements

Performance metrics for each disease class and healthy leaf were evaluated using K-fold cross-validation (K=10). The results are summarized in the tables below:

Table III: Performance Measurements of Leaf Blast

Algorithm	F1-score	Precision	Recall	support
Random Forest	0.98	0.96	0.97	268
SVM (SVC)	0.97	1.00	0.99	268

Table IV: Performance Measurements of Leaf Blight

Algorithm	F1-score	Precision	Recall	support
Random Forest	0.96	0.97	0.96	341
SVM (SVC)	0.97	0.99	0.98	341

Table V: Performance Measurements of Brown SpotA

Algorithm	F1-score	Precision	Recall	Support
Random Forest	0.97	0.97	0.97	317
SVM (SVC)	0.96	0.98	0.97	317

Table VI: Performance Measurements of Tungro

Algorithm	SVM (SVC)			F1-score
RandomForest				0.97

0.97	Precision	Support
	Recall	261
	0.95	261
	0.92	
	0.96	
	0.97	

5. CONCLUSION

This research has effectively showcased the capability of several machine learning algorithms in classifying paddy leaf diseases using a dataset carefully labeled by agricultural experts. By utilizing Decision Tree, Random Forest, Linear Regression, and Support Vector Machine (SVM) classifiers, the study yielded promising outcomes. Among these, the Random Forest classifier excelled, achieving an outstanding accuracy of 97.22% after a thorough evaluation with K-fold cross-validation. This high accuracy highlights the significant potential of machine learning in advancing agricultural practices through more precise disease detection methods.

6. REFERENCES

[1] K.S. Archana and A. Sahayadhas, "Automatic rice leaf disease segmentation using image processing Techniques," Int. J. Eng. Technol, vol. 7, no. 3.27, 2018, pp. 182-185.

[2] M.A. Islam, M.N.R Shuvo, M. Shamsojjaman, S. Hasan, M.S. Hossain, and T. Khatun, "An automated convolutional neural network-based approach for paddy leaf disease detection," International Journal of Advanced Computer Science and Applications, vol. 12, no. 1, 2021.
[3] M.E. Pothen and M.L Pai, "Detection of rice leaf diseases using image processing," In Proceedings of the 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), 2020, pp. 424-430.

[4] R.P. Narmadha and G. Arulvadivu, "Detection and measurement of paddy leaf disease symptoms using image processing," In Proceedings of the 2017 International Conference on Computer Communication and Informatics (ICCCI), 2017, pp. 1-4.

[5] S. Pavithra, A. Priyadharshini, V. Praveena, and T. Monika, "Paddy leaf disease detection using SVM Classifier," International Journal of Communication and Computer Technologies, vol. 3, no. 1, 2015, pp. 16-20.

[6] Zhang N, Wang M & Wang N, "Precision agriculture-a worldwideoverview", Comput. Electron. Agric., Vol.36, No.2–3, (2002),pp.113–132.

[7] Strange RN & Scott PR, "Plant disease: a threat to global foodsecurity", Phytopathology, Vol.43, (2005), pp.83–116.

[8] Barbedo JG, "Digital image processing techniques for detecting,quantifying and classifying plant diseases", Springerplus, Vol.2,No.1, (2013), pp.660–671.

[9] Phadikar S, Sil J & Das AK, "Rice diseases classification usingfeature selection and rule generation techniques", Comput.Electron. Agric., Vol.90, (2013), pp.76–85.

[10] Shrivastava S, Singh SK &Hooda DS, "Soybean plant foliardisease detection using image retrieval approaches", Multimed.Tools Appl., (2016)

[11] Singh V &Misra AK, "Detection of plant leaf diseases using image segmentation and soft computing techniques", Inf. Process. Agric.,(2017).

[12] Clémen A, Verfaille T, Lormel C &Jaloux B, "A new colour visionsystem to quantify automatically foliar discoloration caused byinsect pests feeding on leaf cells", Biosyst. Eng., Vol.133, (2015),pp.128–140.

[13] Barbedo JGA, "A new automatic method for disease symptomsegmentation in digital photographs of plant leaves", Eur. J. PlantPathol., (2016), pp.1–16.

[14] Pydipati R, Burks TF & Lee WS, "Identification of citrus diseaseusingcolor texture features and discriminant analysis", Comput.Electron. Agric., Vol. 52, No.1–2, (2006), pp.49–59.

[15] Khalid S, Khalil T &Nasreen S, "A survey of feature selection andfeature extraction techniques in machine learning", Sci. Inf. Conf.,(2014), pp.372–378.

[16] Anthonys G &Wickramarachchi N, "An image recognition systemfor crop disease identification of paddy fields in Sri Lanka", ICIIS4th Int. Conf. Ind. Inf. Syst., Conf. Proc., (2009), pp.403–407.

[17] Asfarian A, Herdiyeni Y, RaufA&Mutaqin KH, "Paddy diseasesidentification with texture analysis using fractal descriptors basedonfourier spectrum", Proceeding Int. Conf. Comput. Control.Informatics Its Appl. "Recent Challenges Comput.ControlInformatics", (2013), pp.77–81.

[18] Hamuda E, Ginley BM, Glavin M & Jones E, "Automatic cropdetection under field conditions using the HSV colour space andmorphological operations", Comput. Electron. Agric., Vol.133,(2017), pp.97–107.

[19] Zhang M &Meng Q, "Automatic citrus canker detection from leafimages captured in field", Pattern Recognit. Lett., (2011).

[20] Shrivastava S, Singh SK &Hooda DS, "Color sensing and imageprocessing-based automatic soybean plant foliar disease severitydetection and estimation", Multimed. Tools Appl., Vol.74, No.24,(2015), pp.11467–11484.

*[21] Bai XD, Cao ZG, Wang Y, Yu ZH, Zhang XF & L CN i, "Cropsegmentation from images by morphology modeling in the CIEL*a*b* color space", Comput. Electron. Agric., Vol.99, (2013),pp.21–34.*

[22] Medeiros RS, Scharcanski J & Wong A, "Image segmentation viamulti-scale stochastic regional texture appearance models",Comput. Vis. Image Underst., Vol.142, (2016), pp.23–36.

[23] B Kassimbekova, G Tulekova, V Korvyakov (2018). Problems ofdevelopment of aesthetic culture at teenagers by means of theKazakh decorative and applied arts.Opción, Año 33. 170-186.