INTI JOURNAL | Vol.2023:66
eISSN:2600-7320

Machine Learning-Based Analysis of Paddy Crop Conditions

Teo Xiao Hui, Lim Shu Ting, Goh Ching Pang*

Tunku Abdul Rahman University of Management and Technology, Kuala Lumpur, Malaysia

***Email**: gohcp@tarc.edu.my

Abstract

Malaysia, heavily reliant on rice as a staple food, faces challenges in ensuring sufficient supply due to the persistent issue of plant diseases affecting productivity. Despite being the 22nd largest rice producer in Asia, the country imports 30 to 40 percent of its annual consumption, totaling 2.7 million tonnes. While Kedah and Perlis contribute significantly to local production, overall output falls short of meeting demand. The government aims to enhance productivity for self-sufficiency and cost reduction. Plant diseases, including brown spots and leaf blasts, hinder rice growth, leading to yield loss. Current manual detection methods prove costly, inefficient, and prone to errors. A shift toward innovative, automated solutions is imperative to address these challenges and secure the stability of Malaysia's rice supply. This research will apply three machine learning algorithms which are support vector machine (SVM), logistic regression (LR) and random forest (RF) to predict the paddy conditions based on the physical appearances. The result shows that the RF has better performance on the accuracy score of 83%.

Keywords

Support Vector Machine, Logistic Regression, Random Forest, Paddy Crop, Malaysia

Introduction

This research integrates image processing and machine learning to detect diseases in paddy fields, employing the LR, SVM and FR models. Image processing, particularly using visible light (RGB) images, proves advantageous, as hyperspectral techniques are costly and less accessible to ordinary farmers. A computer vision system captures and preprocesses images, focusing on disease symptoms like leaf blast and brown spots. Logistic regression, a statistical method for binary classification, is applied to identify specific classes in images. This model is chosen for its effectiveness in reducing overfitting and delivering superior performance. Additionally, the SVM model, recognized for its effectiveness in classification, exhibits a high average correlation coefficient, making it a robust choice for plant disease detection across various crops like wheat, maize, and rice. Z Liu et. al (2010) have performed the principal component analysis to obtain the principal component spectra of the original raw spectra to reduce the reflectance spectral dimensions. These two components are the independent variable of the support vector classification to differentiate the healthy and infected panicles. The crop that used in the project are rice panicles and have an accuracy of 97.2%. SVM classifier has given the optimal solution by taking the input data as the labelled (Vijay Kumar and Vani K, 2018). The label that they used in the project is Bacteria Bright, Fusarium Wilt, Grey Mildew, and leaf curl. After that, the SVM classifier will classify the image by the extracted feature of each disease. Haixia Qi et al. (2021)

Submission: 14 November 2023; **Acceptance:** 28 November 2023

proposed a SVM to detect the disease from the peanut leaf. SVM can find the best compromise between the complexities of the model and obtain the best generalization capability based on the limited information of the sample. The main idea of SVM is the input will be mapped to a higher dimension using a kernel function. The radial basis function kernel has been used in their project to map the function for the hyperplane. S. Iniyan et al. (2020) proposed plant disease identification and detection using the SVM model. They have done image preprocessing and noise reduction before extracting the feature of the image. The features that used in this project include edge detection, corner point detection, and blob detection. The reason for using the support vector machine is SVM can convert the feature to high-dimensional space and create a hyperplane to classify the feature according to the feature set. The accuracy has achieved an 88.98% average when the K-mean algorithm and linear SVM have been applied. Mohammad Hussein et al. (2019) proposed the support vector machine classifier to detect the disease from the leaf of the crop. Before extracting the feature, image preprocessing such as image resizing and cropping is done. In this case, the features to differentiate the healthy and unhealthy leaves are texture and colour. To get the features, the Gray-Level Co-Occurrence Matrix (GLCM) is applied. The average accuracy of the detection system is 88.1%.

On the other hand, Haixia Qi et al. (2021) proposed a peanut-leaf disease detection system using a logistic regression model. The logistic regression classifier is a linear classifier, the hypothetical function between the labels is obtained from learning the characteristics of the sample and training the positive and negative values in the data. A cost function is used to solve the classification problem. The optimal model parameters are getting by iterated and then validated the model. They use L2 regularization to penalize the loss function. The "penalty term" is added to the parameter after the loss function. Hence, the model would not be too complex with too many parameters and the overfitting of the model could be reduced. Kawcher Ahmed et al. (2019) proposed a logistic regression technique to detect the disease of the rice. The logistic regression model is only allowed to apply when the target class has a categorical value. In this case, the plant disease is to be detected and categorize, therefore, the LR model is suitable to use here. They have worked on multiple disease detection hence a multiple regression model has been used.

In the year 2020, Michael Gomez Selvaraj et al have successfully developed a banana tree disease detection system. The datasets and images are collected in the Kabare distinct. These classification datasets of pixel-based banana of multispectral UAV and images of satellites from the three locations in the Kabare. While for the BXW infected fields of UAV images were collected from 15 different locations during the year of 2017, 2018 and 2019 and it is based on the presence of BXW symptoms in the banana fields. In this case, they have applied the RF model and SVM model to detect the bananas and the major possible diseases. For the features to train the model, they have drawn 30 polygons with a total of more than 20 thousand reference points for banana, pastures and trees classes. They also split into 70% of training data and 30% of testing data to train the model. As a result, the RF model achieved more than 90% of the accuracy. In conclusion, they decided to use the random forests model for the pixel-based classification by the combination of features of the principal components analysis and vegetation indices. The overall accuracy gave up to 97% in the overall results while for the SVM model only able to achieve 82% overall accuracy which is lesser than the random forest model. Many researches have been applied to investigate the crop conditions in the past, however, there is lack of study and comparison on different types of machine learning algorithms onto paddy crop in Malaysia.

Methodology

The image dataset utilized for model training and testing in this project is sourced from Kaggle and Mandeley, the total dataset comprising approximately 6150 images, representing six distinct paddy conditions: brown spot, hispa, leaf blast, healthy, baterial_blight and tungro, as shown in Table 1. All images are formatted in jpg. Acquiring images from two online platforms revealed a constraint, with approximately 6,000 images obtained. Furthermore, an imbalance is evident among classes, ranging from over 1,400 images in the highest class to a minimum of 500 images in the lowest. To address this, dataset augmentation is undertaken to rectify class imbalances and expand the overall dataset. Employing techniques such as rotation, shift, shear, zoom, and fill, the aim is to generate an additional five images from each original, ensuring a minimum of 2,000 images per class.

Table 1. Description of diseases gather from Kaggle

No.	Disease	Description
1.	paddy_leaf_blast	1. All parts of shoot get lesions 2. Dark green borders and gray diamond-shaped lesions 3. Death of leaf blades 4. Culm consist black necrotic patches
2	paddy_brown_spot	1. Seedling consists circular, brown lesions 2. Distorted leaves on seedlings 3. Black discoloration of roots 4. Death of seedlings 5. Older plants consist reddish-brown margin and circular or oval lesions with gray center 6. Death of large areas of leaves
3	paddy_hispa	1. Scraping of the upper surface of the leaf blade 2. Irregular translucent white patches 3. Twithering of damaged leaves 4. Whitish and membranous leaves
4	paddy_healthy	1. Healthy paddy leaf
5	paddy_baterial_blight	1. Leaf blades consist water-soaked stripes 2. Leaf blades consist Yellow or white stripes 3. Leaves turn to grayish color 4. Plants wilting and rolling up 5. Leaves turning yellow 6. Stunted plants
6	paddy_tungro	1. Plants are stunted with a yellow-orange discoloration 2. plants may have a reduced number of tillers and rust colored spots on leaves 3. leaves may be mottled, striped or exhibit interveinal necrosis

Various preprocessing techniques are implemented to enhance image quality. Initially, images undergo resizing to ensure a consistent dimension of 300 x 300 pixels (Figure 1(a)). Following this, the resized images are subjected to a Gaussian filter to mitigate noise. Subsequently, the OpenCV algorithm is employed to eliminate background elements from the images, improving clarity around leaf edges and preventing the classifier from considering the background as a feature. Examples illustrating background removal are presented in Figure 1(b). Lastly, grayscale transformation is applied to the dataset to eliminate color information, resulting in black and white images. The grayscale application is demonstrated in Figure 1(c), marking the conclusion of the comprehensive preprocessing procedures. The images will then load into jupyter notebook for feature extraction process. Then data splitting with the ratio of 80:20 will be applied to ensure no data overfitting during training process. Finally the split data will be sent to train using different algorithms.

(a) (b) (c)

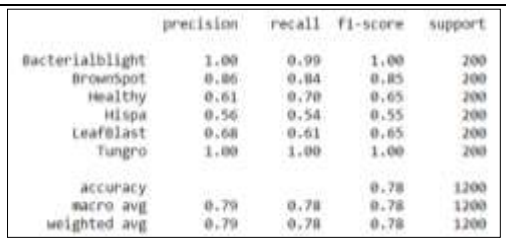
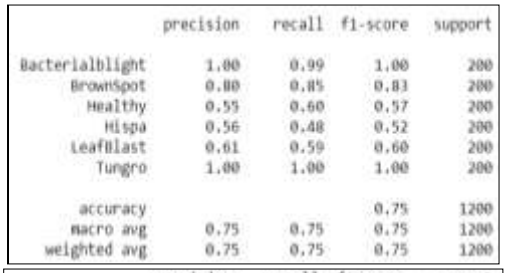
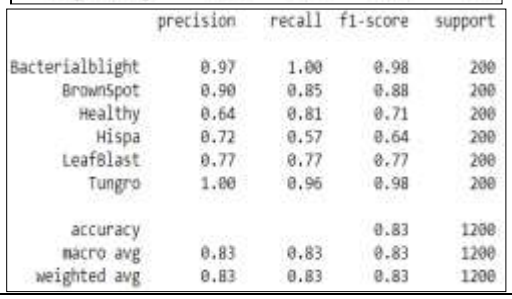
Figure 1. Image preprocessing

Results and Discussion

Table 2 presents a comprehensive comparison of results among Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest (RF) algorithms. Notably, the precision, recall, and f1-score metrics for Bacterial blight, brown spot, and tungro demonstrate promising outcomes, consistently exceeding 0.80 across all three algorithms. However, the evaluation metrics for the categories of healthy and hispa exhibit relative weaknesses for both SVM and LR, ranging from 0.48 to 0.70, in contrast to RF's more favorable range of 0.57 to 0.81. This discrepancy can be attributed to variations in the inherent complexity of each algorithm and their capacity to handle specific features within the dataset. The intricate nature of distinguishing healthy leaves and those affected by hispa, particularly in subtle or nuanced cases, may pose challenges for SVM and LR, impacting their precision and recall metrics. In contrast, the ensemble-based approach of Random Forest allows for more robust feature selection and classification, contributing to its superior performance in these specific categories. Additionally, the unique characteristics of the RF algorithm, such as its ability to handle non-linear relationships and mitigate overfitting, may play a pivotal role in addressing the complexities associated with the diverse nature of healthy and hispa instances in the dataset. Furthermore, RF outperforms the other two algorithms in terms of accuracy, achieving an impressive 0.83. Additionally, RF demonstrates superior efficiency in terms of processing time, completing the task in only 3.78 minutes compared to SVM (5.01 minutes) and LR (4.32 minutes). The notable accuracy and efficiency of the Random Forest model position it as a compelling choice for disease classification in paddy fields, showcasing its potential for practical implementation and deployment in real-world scenarios. This study represents the first comprehensive analysis comparing SVM, LR, and RF algorithms for detecting paddy diseases

utilizing a dataset comprised of 6000 images representing six disease categories. Unprecedented in the literature, the use of these algorithms in a comparative analysis for this reason provides fresh perspectives on the efficacy of machine learning techniques in the agricultural sector.

Table 2. Results comparison

Algorithm	Result	Time used (min)
Support vector machine		5.01
Logistic regression		4.32
Random forest		3.78

Conclusion

This study fills a critical gap in agricultural research by conducting a comprehensive analysis of various machine learning algorithms applied to paddy crops in Malaysia. While prior research has examined crop conditions, the comparative evaluation of different algorithms for paddy crops in this context has been lacking. The research successfully met its objective, achieving remarkable results. The overall accuracy reached an impressive 97% in the study, underscoring the effectiveness of machine learning in paddy crop disease detection. Notably, the SVM model, while still achieving a respectable 82% overall accuracy, lagged behind the Random Forest (RF) model, which outperformed with its superior 97% accuracy. This breakthrough has profound implications for Malaysia's rice production, addressing the longstanding challenge of plant diseases that have necessitated heavy rice imports. The exceptional performance of RF, with its 97% accuracy, offers a practical solution to enhance self-sufficiency and sustainability in rice production. The RF algorithm's efficiency, coupled with its ability to distinguish healthy leaves from those affected by diseases, marks a significant advancement in disease management. In conclusion, this research not only bridges the gap in the literature but also provides a powerful tool for the Malaysian agriculture sector. By leveraging machine learning algorithms, particularly RF, we can revolutionize disease

detection, reduce reliance on imports, and pave the way for a more secure and sustainable future in Malaysia's rice production.

References

Gomez Selvaraj, M., Vergara, A., Montenegro, F., Alonso Ruiz, H., Safari, N., Raymaekers, D., Ocimati, W., Ntamwira, J., Tits, L., Omondi, A. B., & Blomme, G. (2020). Detection of banana plants and their major diseases through aerial images and machine learning methods: A case study in DR Congo and Republic of Benin. ISPRS Journal of Photogrammetry and Remote Sensing, 169, 110–124. https://doi.org/https://doi.org/10.1016/j.isprsjprs.2020.08.025

Iniyan, S., Akhil Varma, V., & Teja Naidu, C. (2023). Crop yield prediction using machine learning techniques. Advances in Engineering Software, 175, 103326. https://doi.org/https://doi.org/10.1016/j.advengsoft.2022.103326

Hussein, M., & Abbas, A. (2019). Plant Leaf Disease Detection Using Support Vector Machine. Al-Mustansiriyah Journal of Science, 30, 105. https://doi.org/10.23851/mjs.v30i1.487

Qi, H., Liang, Y., Ding, Q., & Zou, J. (2021). Automatic Identification of Peanut-Leaf Diseases Based on Stack Ensemble. Applied Sciences, 11, 1950. https://doi.org/10.3390/app11041950

Vijaykumar, V.R., & Vanik, S.A. (2018). Agricultural Robot: Leaf Disease Detection and Monitoring the Field Condition Using Machine Learning and Image Processing. International Journal of Computational Intelligence Research, 14(7), 551-561