



CLASSIFICATION OF DISEASES AND PESTS OF MAIZE USING MULTINOMIAL LOGISTIC REGRESSION BASED ON RESAMPLING TECHNIQUE OF K-FOLD CROSS-VALIDATION

YULIA RESTI*, DESI HERLINA SARASWATI, DES A. ZAYANTI, NING ELIYATI

Departement of Mathematics, Faculty of Mathematics and Natural Science, Universitas Sriwijaya, Inderalaya, Indonesia

**Corresponding author: yulia_resti@mipa.unsri.ac.id*

(Received: 08 August 2022; Accepted: 13 September 2022; Published online: 01 November 2022)

ABSTRACT: Some of the obstacles in the cultivation of maize that cause low productivity of maize yields are diseases and pests. Early detection of maize diseases and pests is expected to reduce farmer losses. A system for the early detection of diseases and pests can be created by classifying them based on digital images. This study aimed to classify maize diseases and pests using multinomial logistic regression. The model and testing resampling were based on the resampling technique of k-fold cross-validation. The research data was obtained from the RGB color feature extraction process for each object in each class of diseases and pests of corn. The results showed that the classification into seven classes using five folds had an accuracy rate of 99.85%, macro precision of 98.59%, and macro recall of 98.15%.

KEYWORDS: *Classification, Multinomial Logistic Regression, and Repeated k-Fold Cross Validation.*

1. INTRODUCTION

Maize is a food crop used as the primary raw material for industry and animal feed [1]. Diseases and pests are the causes of low maize yields and can even cause crop failure. If pests attack corn, it can result in a 70% loss of crop yields. Disease attacks on corn can cause yield losses of up to 90%, while pest attacks can experience crop losses of up to 70% [2].

Multinomial logistic regression is a statistical analysis method used to determine the relationship between one or more predictor variables and response variables with categorical data types and more than two categories [3]. Multinomial logistic regression can also be used for classification tasks in machine learning [4]. This method classifies the obtained probability-based data from the model [5], and the application in several classification cases gives various performances. For example, the classification of Covid-19 patients provides accuracy and sensitivity of 98.2% and 98.8%, respectively [6]. Newborn weight classification also provides accuracy above 90% by 91.8% [7]. Classification of the difficulty level of learning statistics courses provides an accuracy of 76.5% [8]. The classification of the credit scoring of Portuguese financial institutions provides an accuracy of up to 89.79% [9].

In the last decade, the use of digital images for object classification has been prevalent because of its low cost [10], especially for diseases and pests of plants [11]–[16]. Classification of diseases and pests of maize based on digital images can be used to build an early detection system [17], [18]. This system helps reduce the opportunity for farmers to lose due to low



productivity or crop failure caused by diseases and pests [19], [20]. This study aimed to classify maize diseases and pests using multinomial logistic regression. The model and testing resampling were based on k -fold cross-validation. This technique is recommended because the classification performance obtained is more accurate since it produces multiple datasets so that the model's performance is calculated as the average of the k datasets [5].

2. RESEARCH METHOD

2.1. Data

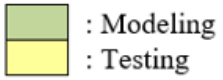
Digital images of the diseases and pests that affect corn plants serve as the study's data source. The digital images were collected between September and October 2021 using a 12-megapixel smartphone camera located in corn plantations of the villages of Tanjung Pering, Tanjung Seteko, and Tanjung Baru, Ogan Ilir Regency of South Sumatra, Indonesia, where the captures took place. These areas are close to the University of Sriwijaya. The data of 4616 digital images consisting of one class of NP (non-pathogen) at 23.2%, three classes of diseased at 42.89%, and three classes of pest-infested maize plants at 33.90% were distributed as a consequence of this inquiry. Three classes of diseases are LRD (leaf rust disease) at 29.96%, DWD (downy mildew disease) at 1.06%, and LBD (leaf blight disease) at 11.87%. Three classes of pests are LP (Locusta pests) at 2.34%, SFP (Spodoptera Frugiperda pest) at 28.96%, and HAP (Heliotis Armigera pest) at 2.6%.

2.2. Methodology

The steps of this research are given below:

1. The preprocessing data. This stage consists of 2 processes, the first is to crop the image, and the second is to extract the red, green, and blue (RGB) color features.
2. Split data randomly into k -fold for $k = 5$, as shown in Fig. 1.

Composition	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
1					
2					
3					
4					
5					



: Modeling
: Testing

Fig. 1. The k -fold cross-validation resampling technique for $k = 5$

3. Modeling the first $(k - 1)$ fold data using multinomial logistic regression
4. Testing using 1-fold the remaining data
5. Repeat the third and fourth steps for each other composition so that there are k sets of models that must be evaluated.
6. Determine the model performance, which is the average of the k sets of models. The model performance is calculated based on the multiclass confusion matrix (Table 1), which consists of accuracy rate, macro precision, and macro recall.



Table 1: Confusion matrix for the first class of diseases and pests of the corn plant

		Prediction Class						
		<i>j</i>	LRD	DWD	LBD	LP	SFP	HAP
Actual class	LRD		TP	FN	FN	FN	FN	FN
	DWD		FP	TN	TN	TN	TN	TN
	LBD		FP	TN	TN	TN	TN	TN
	LP		FP	TN	TN	TN	TN	TN
	SFP		FP	TN	TN	TN	TN	TN
	HAP		FP	TN	TN	TN	TN	TN

source: Resti et al., 2022a

For modeling using multinomial logistic regression, the parameter can be estimated using Maximum Likelihood Estimation and Newton-Raphson. The likelihood function obtained by observations that assumed each pair of observations to be independent,

$$l(\beta) = \prod_{i=1}^n f(x_i) = \prod_{i=1}^n \pi_1(x_i)^{y_{1i}} \pi_2(x_i)^{y_{2i}} \dots \pi_j(x_i)^{y_{ji}} \quad (1)$$

The concept of Maximum Likelihood Estimation (MLE) states the estimated value β that maximizes the likelihood function, which is the solution to the first derivative of the likelihood function,

$$L(\beta) = \ln[l(\beta)]$$

$$L(\beta) = \sum_{i=1}^n y_{1i} \ln[\pi_1(x_i)] + y_{2i} \ln[\pi_2(x_i)] + \dots + y_{ji} \ln[\pi_j(x_i)] \quad (2)$$

Furthermore, an explicit solution is obtained using the Newton-Raphson method of the second differential $L(\beta)$ to β ,

$$\frac{\partial^2 L(\beta)}{\partial \beta_{jk} \partial \beta_{jk'}} = \sum_{i=1}^n x_{1i} x_{2i} \dots x_{ji} \pi_1(x_i) \pi_2(x_i) \dots \pi_j(x_i) \quad (3)$$

The statistics for simultaneous testing whether the independent variable as a whole has a significant effect on the dependent/target variable is written as,

$$G = -2 \ln \left[\frac{l_0}{l_k} \right] \quad (4)$$

For l_0 be the likelihood without an independent variable and l_k be the likelihood with independent variables with $k = 1, 2, \dots, p$. The null hypothesis is no independent variable that is statistically significant in influencing the dependent variable ($\beta_1 = \beta_2 = \dots = \beta_p = 0$), and the alternative is at least one independent variable that statistically significantly affects the dependent variable (at least one $\beta_k \neq 0$). Reject criteria for the null hypothesis when $G > X_{a,df}^2$, where the degree of freedom (df) is the number of independent variables.

The Wald statistics for partial test whether the independent variable affects the dependent variable in individuals written as



$$W_k = \left(\frac{\hat{\beta}_k}{SE(\hat{\beta}_k)} \right)^2 \quad (5)$$

For $\hat{\beta}_k$ be the estimator of β_k and $SE(\hat{\beta}_k)$ be the standard error of $\hat{\beta}_k$ with $k = 1, 2, \dots, p$. The null hypothesis is no independent variable that partially affects the dependent variable ($\beta_k = 0$), and the alternative is an independent variable that partially affects the dependent variable ($\beta_k \neq 0$). The criteria for rejecting the null hypothesis $p - value < \alpha$, with α being the significance level.

The statistics for the goodness of fit test were given as follows,

$$Deviance = -2 \sum_{i=1}^n \sum_{j=1}^J y_{ij} \log \frac{y_{ij}}{\hat{\mu}_{ij}} \quad (6)$$

For $\hat{\mu}_{ij} = n_i \hat{\pi}_{ij}$ and J be the number of classes in the response variable. The null hypothesis is the appropriate model for use. The criterion for rejection of the null hypothesis is $Deviance > \chi^2_{(\alpha; (n-p)(r-1))}$ or $p - value < \alpha$.

3. RESULT AND DISCUSSION

Image cropping focuses on the part of the leaves with diseases or pests. The example of the image result after cropping is presented in Fig. 2.

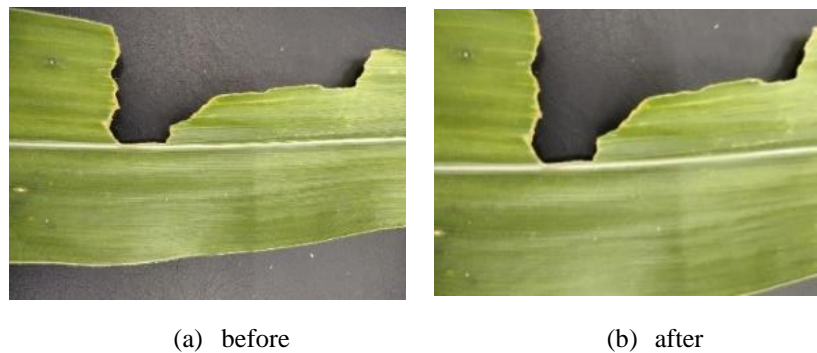


Fig. 2. The image before and after cropping

The mean of feature extraction of R, G, and B variables is given in Table 2. The HAP class has the highest mean value for the R feature, and LRD class has the lowest mean value compared to the other classes. For G and B features, respectively, the highest mean value belongs to the DWD and Healthy classes, while the lowest mean value belongs to the LRD class.

Table 3 gives a simultaneous test to determine whether the independent variables have a significant simultaneous effect on the dependent variable. The G value that is greater than the $X^2_{(0,05;18)}$ indicates that the null hypothesis is rejected. All of the image features affect the classes of maize diseases and pests.



Table 2: Mean of independent variables

Class	R	G	B
NP	128.45	163.34	121.61
LBD	119.13	125.56	93.11
DWD	148.20	163.86	79.79
LRD	97.99	105.00	77.80
HAP	163.23	149.13	89.37
SFP	120.57	150.02	64.53
LP	139.75	152.90	96.21

Table 3: Simultaneous test

$-2\ln(l_0)$	$-2\ln(l_k)$	G	$\chi^2_{(0.05;18)}$	p-value
11356.41	3531.49	7824.92	28,87	0.00

Table 4: Goodness of fit test

Deviance	df	p-value
3531.49	22074	1.00

Table 5: Parameter estimation

Diseases & Pests Class		β	SE(β)	Wald	p-value	exp(β)
NP (non patogen)	Intercept	-10.70	1.79	35.60	0.00	
	R	-0.82	0.04	388.18	0.00	0.44
	G	0.62	0.04	317.83	0.00	1.85
	B	0.23	0.01	285.78	0.00	1.26
LBD (leaf blight disease)	Intercept	18.25	1.53	141.88	0.00	
	R	0.0	0.02	0.73	0.39	1.02
	G	-0.18	0.02	96.94	0.00	0.83
	B	0.08	0.01	66.73	0.00	1.08
DWD (downy mildew disease)	Intercept	-11.71	2.53	21.40	0.00	
	R	0.06	0.02	6.78	0.01	1.06
	G	0.04	0.02	3.50	0.06	1.04
	B	-0.04	0.02	24.00	0.00	0.96
LRD (leaf rust disease)	Intercept	26.53	1.58	282.71	0.00	
	R	-0.02	0.02	1.38	0.24	0.98
	G	-0.24	0.02	147.30	0.00	0.79
	B	0.11	0.01	123.10	0.00	1.12
HAP (Heliotis Armigera pest)	Intercept	1.52	2.56	0.35	0.55	
	R	0.19	0.02	77.74	0.00	1.21
	G	-0.18	0.02	54.42	0.00	0.84
	B	-0.04	0.01	8.39	0.00	0.96
SFP (Spodoptera Frugiperda pest)	Intercept	3.54	1.41	6.32	0.01	
	R	-0.23	0.02	148.34	0.00	0.80
	G	0.18	0.02	104.30	0.00	1.20
	B	0.02	.006	7.66	0.01	1.02

The goodness of fit test using deviance, as presented in Table 4, informs that the null hypothesis cannot be rejected because the *p-value* is greater than the significance level. The model obtained is feasible to use with a 95% confidence level.

The parameter estimation, including partial test and odds ratio, are presented in Table 5. We take the Locusta pest (LP) class as a reference for modeling in all data sets. The table



informs that if a digital image of the corn plant NP class has a pixel value of variable R that increases by one, the pixel value of the other variables remains constant. The digital image's relative risk (tendency) to be classified as NP class will decrease by 0.82 pixels. If the pixel value of the G variable in the NP class increases by one and the pixel values of the other variables remain constant, then the digital image's relative risk (tendency) to be classified as NP class will increase by 0.62 pixels. Likewise, the interpretation for the estimated parameters of variable B and the parameters that are significant in other diseases and pest classes.

Table 6 indicates all of the images features in each class of maize disease and pest significantly affect the model except for the red feature in the LBD and the LRD classes and the blue feature in the DWD class. The odds ratio of 0.44 for the R variable in the NP class indicates that a digital image has a smaller tendency (0.44 times) to be classified into the NP class than the LP class based on the R variable. A digital image has a greater tendency (1.85 times) to be classified. The NP class is compared to the LP class based on variable B. Likewise, the interpretation for the odds ratio for variable B and the odds ratio for variables that are significant in other diseases and pest classes. The best model for the first dataset composition of k-fold cross-validation is given in Table 6.

Table 6: The best model of multinomial logistic regression

Class (j)	$\pi_j(x)$
NP	$\frac{\exp(-10.7 - (0.82x_1) + (0.62x_2) + (0.23x_3))}{1 + \exp(-10.7 - (0.82x_1) + (0.62x_2) + (0.23x_3))}$
LBD	$\frac{\exp(18.25 - (0.18x_2) + (0.08x_3))}{1 + \exp(18.25 - (0.18x_2) + (0.08x_3))}$
DWD	$\frac{\exp(-11.71 - (0.06x_1) - (0.04x_3))}{1 + \exp(-11.71 - (0.06x_1) - (0.04x_3))}$
LRD	$\frac{\exp(26.53 - (0.24x_2) + (0.11x_3))}{1 + \exp(26.53 - (0.24x_2) + (0.11x_3))}$
HAP	$\frac{\exp((0.19x_1) - (0.18x_2) - (0.04x_3))}{1 + \exp((0.19x_1) - (0.18x_2) - (0.04x_3))}$
SFP	$\frac{\exp(3.54 - (0.23x_1) + (0.18x_2) + (0.02x_3))}{1 + \exp(3.54 - (0.23x_1) + (0.18x_2) + (0.02x_3))}$

The overall performance of the proposed classification model using multinomial logistic regression is the average performance of the k sets of models for k = 5. Ultimately presented in Table 7.

Table 7: Overall performance of the multinomial logistic regression (percentage)

Dataset	Average Accuracy	Precision _M	Recall _M
1	99.75	97.93	94.73
2	99.91	99.74	99.90
3	99.83	99.06	98.18
4	99.88	98.09	99.29
5	99.89	98.15	98.65
Average	99.85	98.59	98.15
Stdev	0.03	0.78	2.02

The macro precision has the highest value fluctuation, followed by a macro recall and average accuracy successively. The fact that the values are different in each dataset is why the



measurements obtained using the repeated k-fold cross-validation resampling technique has a more accurate value. The overall performance of the multinomial logistic regression for $k = 5$ datasets shows that the proposed multinomial logistic regression model is quite good. The proposed model can correctly classify all digital images into each class of 99.85%. Furthermore, the proposed model can classify digital images in a class that is not a class by 98.59% and classifies digital images that are correct in each class by 98.15%.

4. CONCLUSION

Diseases and pests are barriers to maize production that result in low productivity of maize crops. Reduced farmer losses are anticipated with the early diagnosis of maize diseases and pests. An early detection system of diseases and pests can be developed based on classification types based on digital images. The extraction of RGB color features from each object in each class of maize diseases and pests was used to gather the research data. The implementation of multinomial logistic regression based on the resampling technique of k-fold cross-validation obtained the results that the proposed model has a 99.85% accuracy rate for correctly classifying all digital images into each class. The proposed model also correctly classifies digital images in a class that is not a class by 98.59% and can correctly classify digital images in a class by 98.15%.

ACKNOWLEDGEMENT

This paper is the discussions result with the smart farming development team and the smart inspection system discussion group of the Universitas Sriwijaya.

REFERENCES

- [1] L. Sumaryanti, T. Istanto, and S. Pare, "Rule Based Method in Expert System for Detection Pests and Diseases of Corn," *J. Phys. Conf. Ser.*, vol. 1569, no. 2, 2020, doi: 10.1088/1742-6596/1569/2/022023.
- [2] A. T. Sumpala and R. Rasyid, "Expert system for corn plant disease diagnosis with the breadth-first search method," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 382, no. 1, 2019, doi: 10.1088/1755-1315/382/1/012001.
- [3] A. J. Scott, D. W. Hosmer, and S. Lemeshow, "Applied Logistic Regression.," *Biometrics*, vol. 47, no. 4, p. 1632, 1991, doi: 10.2307/2532419.
- [4] Y. Resti, E. S. Kresnawati, N. R. Dewi, D. A. Zayanti, and N. Eliyati, "Diagnosis of diabetes mellitus in women of reproductive age using the prediction methods of naive bayes, discriminant analysis, and logistic regression," *Sci. Technol. Indones.*, vol. 6, no. 2, pp. 96–104, 2021, doi: 10.26554/STI.2021.6.2.96-104.
- [5] J. Gareth, W. Daniela, H. Trevor, and T. Robert, *An Introduction to Statistical Learning, with Applications in R*, vol. 7, no. 10. New York: Springer New York Heidelberg Dordrecht London, 2013.
- [6] N. Raoo, S. Mostafa, and A. Hadi Kazemi, "Social Support and Self - Care Behavior Study," *J. Educ. Health Promot.*, vol. 11, no. May, pp. 1–6, 2022, doi: 10.4103/jehp.jehp.
- [7] S. P. Sari, I. Suliansyah, N. Nelly, and H. Hamid, "Identifikasi Hama Kutudaun (Hemiptera: Aphididae) Pada Tanaman Jagung Hibrida (*Zea Mays* L.) Di Kabupaten Solok Sumatera Barat," *J. Sains Agro*, vol. 5, no. 2, 2020, doi: 10.36355/jsa.v5i2.466.
- [8] A. Abdillah, A. Sutisna, I. Tarjiah, D. Fitria, and T. Widiyanto, "Application of Multinomial Logistic Regression to analyze learning difficulties in statistics courses," *J. Phys. Conf. Ser.*, vol. 1490, no. 1, 2020, doi: 10.1088/1742-6596/1490/1/012012.
- [9] E. Costa e Silva, I. C. Lopes, A. Correia, and S. Faria, "A logistic regression model for consumer default risk," *J. Appl. Stat.*, vol. 47, no. 13–15, pp. 2879–2894, 2020, doi: 10.1080/02664763.2020.1759030.
- [10] L. C. Ngugi, M. Abelwahab, and A.-Z. Mohammed, "Recent advances in image processing



- techniques for automated leaf pest and disease recognition A review," *Inf. Process. Agric.*, vol. 8, pp. 27–51, 2021, doi: <https://doi.org/10.1016/j.inpa.2020.04.004>.
- [11] T. S. Xian and R. Ngadiran, "Plant Diseases Classification using Machine Learning," in *The 1st International Conference on Engineering and Technology (ICoEngTech) 2021*, 2021, vol. 1962, no. 012024, pp. 1–12, doi: [10.1088/1742-6596/1962/1/012024](https://doi.org/10.1088/1742-6596/1962/1/012024).
- [12] A. K. Singh, B. Chourasia, N. Raghuwanshi, and K. Raju, "BPSO based feature selection for rice plant leaf disease detection with random forest classifier," *Int. J. Eng. Trends Technol.*, vol. 69, no. 4, pp. 34–43, 2021, doi: [10.14445/22315381/IJETT-V69I4P206](https://doi.org/10.14445/22315381/IJETT-V69I4P206).
- [13] P. Sutha, A. Nandhu Kishore, V. Jayanthi, A. Periyanan, and P. Vahima, "Plant Disease Detection Using Fuzzy Classification," *Ann. R.S.C.B.*, vol. 25, no. 4, pp. 9430–9441, 2021.
- [14] M. Syarief and W. Setiawan, "Convolutional neural network for maize leaf disease image classification," *Telkomnika (Telecommunication Comput. Electron. Control.)*, vol. 18, no. 3, pp. 1376–1381, 2020, doi: [10.12928/TELKOMNIKA.v18i3.14840](https://doi.org/10.12928/TELKOMNIKA.v18i3.14840).
- [15] T. Kasinathan, D. Singaraju, and R. U. Srinivasulu, "Insect classification and detection in field crops using modern machine learning techniques," *Inf. Process. Agric.*, vol. 8, no. 3, pp. 446–457, 2021, doi: <https://doi.org/10.1016/j.inpa.2020.09.006>.
- [16] B. Rajesh, M. V. S. Vardhan, and L. Sujihelen, "Leaf Disease Detection and Classification by Decision Tree," in *Machine Learning Foundations*, 2020, no. Icoei, pp. 141–165, doi: [10.1007/978-3-030-65900-4_7](https://doi.org/10.1007/978-3-030-65900-4_7).
- [17] K. P. Panigrahi, H. Das, A. K. Sahoo, and C. S. Moharana, *Maize Leaf Disease Detection and Classification Using Machine Learning Algorithms*, no. January. Springer, Singapore., 2020.
- [18] B. S. Kusumo, A. Heryana, O. Mahendra, and H. F. Pardede, "Machine Learning-based for Automatic Detection of Corn-Plant Diseases Using Image Processing," in *2018 International Conference on Computer, Control, Informatics and its Applications: Recent Challenges in Machine Learning for Computing Applications, IC3INA 2018 - Proceeding*, 2019, pp. 93–97, doi: [10.1109/IC3INA.2018.8629507](https://doi.org/10.1109/IC3INA.2018.8629507).
- [19] Y. Resti, C. Irsan, M. Amini, I. Yani, R. Passarella, and D. A. Zayanti, "Performance Improvement of Decision Tree Model using Fuzzy Membership Function for Classification of Corn Plant Diseases and Pests," *Sci. Technol. Indones.*, vol. 7, no. 3, pp. 284–290, 2022, doi: [10.26554/sti.2022.7.3.284-290](https://doi.org/10.26554/sti.2022.7.3.284-290).
- [20] Y. Resti, C. Irsan, M. T. Putri, I. Yani, Anshori, and B. Suprihatin, "Identification of Corn Plant Diseases and Pests Based on Digital Images using," *Sci. Technol. Indones.*, vol. 7, no. 1, pp. 29–35, 2022, doi: <https://doi.org/10.26554/sti.2022.7.1.29-35>.