https://doi.org/10.17221/25/2024-RAE

Prediction of Physicochemical characteristics of lemon (*Citrus Limon* cv. Montaji Agrihorti) using Vis-NIR spectroscopy and machine learning model

*Jihan Nada Salsabila Erha, Dina Wahyu Indriani, Zaqlul Iqbal, Bambang Susilo, Dimas Firmanda Al Riza**

Department of Biosystems Engineering, Faculty of Agricultural Technology, University of Brawijaya, Malang, Indonesia

**Corresponding author: dimasfirmanda@ub.ac.id*

Citation: Erha J.N.S., Indriani D.W., Iqbal Z., Susilo B., Al Riza D.F. (2024): Prediction of physicochemical characteristics of lemon (*Citrus limon* cv. Montaji Agrihorti) using Vis-NIR spectroscopy and machine learning model. Res. Agr. Eng., 70: 218–225.

Abstract: Lemons are fruit products that grow well in Indonesia. Montaji Agrihorti is one of the lemon varieties found in Indonesia, a new variety developed by Balitjestro breeding. This lemon variety is seedless. In fact, lemons are harvested nearly all year-round. Equally important, evaluating the fruit's maturity level is crucial for determining the optimal harvest time. In this study, standardizing measurement on maturity level was conducted through Vis-NIR spectroscopy and machine learning models. In this case, non-destructive data from Vis-NIR spectroscopy were correlated with parameters related to fruit maturity and quality, such as soluble solid content (SSC), acidity, firmness, essential oil yield, and essential oil content. Non-destructive test involved capturing spectral data to be subsequently processed through machine learning models such as SVM, KNN, and random forest. The most accurate results were obtained using the SVM method for SSC and firmness parameters, with accuracy of 72 and 78%, respectively. For visual and acidity parameters, the most accurate result was performed through random forest with visual accuracy value 94% for all features, all features-MA (moving averages) was 97%, 36-PCA (principal component analysis) was 94%, and 36-PCA-MA was 97%. As for acidity, the accuracy for all features was 89%, all features-MA was 81%, 36-PCA was 89%, and 36-PCA-MA was 83%.

Keywords: destructive; KNN; montaji Agrihorti Lemon; random forest; SVM; chemometrics

Lemon is a fruit product that grows well in Indonesia. Montaji Agrihorti is one of the lemon varieties found in Indonesia, a new variety developed through Balitjestro breeding. According to a previous study, following oranges, lemons are the third significant citrus fruit annually cultivated, accounting for over 4.4 million tons. Lemon, scientifically known as *Citrus limon*, provides significant health benefits due to its richness in vitamin C, potassium, magnesium, minerals, citric acid, and high levels of flavonoids. Lemons contain 3.7% citric acid and approximately 4050 mg·100 g^{-1} of Vitamin C (Kieling et al. 2018; Rafique et al. 2020). Lemons have anticancer properties and exhibit antibacterial activities as a result of alkaloids in various parts of the plant, including leaves, stems, roots, and flowers. Lemon peels, moreover, contain diverse phytochemicals, such as essential oils, glycosides,

Supported by Faculty of Agricultural Technology, University of Brawijaya through Non-Associate Professor Doctoral Research Grant 2022 (DNLK) (Contract No.: 12938/UN10/TU/2022).

https://doi.org/10.17221/25/2024-RAE

and β and γ-sitosterol (Nurlatipah et al. 2017). One of the lemon varieties in Indonesia, Montaji Agrihorti, is a new variety bred by the Indonesian Citrus and Subtropical Fruit Research Institute (Balitjestro). In this case, the Indonesian government exhibited Montaji lemon as a new variety in 2018 under Minister of Agriculture of Indonesia (Decree No. 039/Kpts/SR.120/D.2.7.4.2018). Moreover, the lemon variety provides health benefits due to its thin skin and high juice composition. Further, Montaji Agrihorti is seedless. In addition, this lemon variety is harvested nearly annually.

Fruit harvesting process involves sorting based on size and surface defects. Additionally, measuring the fruit quality, such as fruit ripeness, is performed by observing harvest characteristics, including skin color and texture, based on cultivation experience (Bhargava and Bansal 2021). However, evaluating lemon quality is commonly conducted through conventional practices. Measuring and determining the maturity level is conducted visually, while measuring and determining the maturity level are conducted through the extraction process, which is subsequently followed by a laboratory test involving manual tests on total dissolved solids, sugar level, and vitamins. Nevertheless, the techniques involve destructive practices; thus, a non-destructive technique is required to measure the fruit quality. In general, the quality measurement process is conducted manually, leading to less accuracy and objective results due to a lack of human capacity. Moreover, destructive testing requires an extended period of time or unable to be conducted real-time as the test involves diverse stages of processing. Therefore, developing a more effective process is required to predict the maturity level of lemons by utilising non-destructive alternative technologies. Furthermore, non-destructive prediction of fruit ripeness has been studied as an innovative approach for fruit selection.

Non-destructive optical method based on Vis-NIR (visible – near infra-red) represents an approach that allows fruit analysis that excludes chemicals, in addition, is environmentally friendly, and generates high accuracy. The Nis-VIR spectroscopy mechanism involves fruit samples that are exposed to light from a spectrometer. The light is reflected back, thus generating spectral data, in which the reflected light carries information on the internal composition of the sample (Lim et al. 2014). Vis-NIR technology is widely practiced for non-destructive quality prediction in agricultural commodities. In this case, implementing Vis-NIR spectroscopy requires an analysis to extract information from spectral data (Hadiwijaya et al. 2020).

Data analysis utilised as an alternative measurement is the support vector machine (SVM) that employs hypotheses on linear functions within high dimensional features (Zhao et al. 2013), in addition to K-Nearest Neighbors algorithm method (KNN), and random forest. Further, the objective of utilising the methods is to discover the most accurate prediction. In practice, the measurement was conducted using machine learning model, an artificial intelligence, allowing software applications to predict the results more accurately.

Following the constraints, the findings of the study are expected to create highly accurate predictive modelling of physicochemical parameters for lemon fruit for further development, thus facilitating a better harvesting of Montaji Agrihorti lemon. Additionally, developing this predictive model generates a correlation between the characteristics of Vis-NIR spectroscopy of lemons and the results of destructive measurements at different maturity levels.

MATERIAL AND METHODS

Material and equipment. The study was conducted on Montaji Agrihorti lemons as samples through three levels of visual maturity (unripe, half-ripe, and ripe), with a total of 120 samples. The study utilised a Vis-NIR spectroscopy setup of different components, including the Go Direct® SpectroVis® Plus Spectrophotometer (Vernier, USA), optical cables, halogen lamp, IR lamp, laptop, and

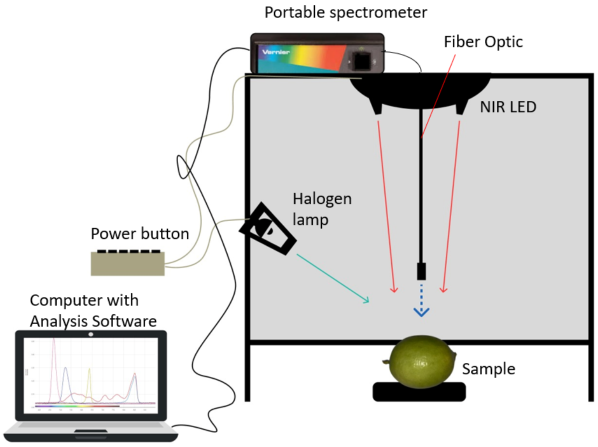

Figure 1. Non-destructive measurement experimental setup

software spectral analysis (Figure 1). Destructive data were collected through penetrometers to measure fruit firmness level, pocket brix acidity portable for total soluble solids and fruit acidity. In addition, a simple distillation set up to extract essential oil from the lemon peels.

Research methods. This study involved collecting spectral data on lemon fruits using the diffuse reflectance technique with halogen and infrared lamps. The process applied visible light (Vis; 400–750 nm) and Near Infrared (NIR; 750–2500 nm). Samples for Montaji Agrihorti were taken through visual grouping into three different maturity levels: unripe (green), half-ripe (greenish-yellow), and ripe (yellow). 40 fruits were selected for each maturity level, with total samples of 120 fruits. On top of this, physical and chemical tests or destructive data acquisition was performed in addition to collecting spectral data, including fruit firmness test, measurement of Total Soluble Solids presented in °brix, acidity, and essential oil extraction from Montaji Agrihorti peels.

Data collection mechanism for Montaji Agrihorti lemons started from preparing the samples, which involved selecting cleaned Montaji Agrihorti and, afterwards, taking non-destructive data using spectral reflectance measurement. Next, proceed with destructive methods including the fruit firmness test, brix, acidity, and essential oil distillation. Furthermore, perform a GC-MS test on the distilled essential oil.

RESULTS AND DISCUSSION

Spectral data collection was performed destructively using 120 Montaji lemons, in which 40 fruits were taken from each maturity level. The data were then grouped into three indexes to facilitate data processing. Grouping or changing the data form was performed by inputting the available destructive data by providing a specific range for each maturity parameter. The range grouping was created by determining the maximum limit, median, and minimum values for each set of data, which were subsequently processed in Microsoft Excel (version 2021). In this case, data class calculation was performed using the Sturges method. The Sturges method is a rule applied to determine the optimal number of bins for a dataset (Dogan et al. 2010). In this particular context, the objective of determining the class for each parameter is to facilitate data processing. On the other hand, the non-destructive spectral data collection generated data in the form of .csv file consisting of wavelength and intensity data.

The relationship of maturity level (visual) and fruit firmness. Measuring fruit firmness was performed using a fruit penetrometer with 120 samples and a total of 40 fruits. Samples for each maturity level of Montaji lemon fruit is presented in Figure 2, in which each maturity level acquired a different range of values. Figure 3 presents boxplot of firmness data, which show decreasing trend of firmness during ripening. This result is in line with the previous study, where the decrease in fruit firmness is attributed to changes in the composition of cell walls, resulting in fruit softening (Samaradiwakara et al. 2019).

The relationship between maturity level and total soluble solids. Measuring fruit firmness was performed on three maturity levels, in which the Y-axis, as presented in the figure, is the range of total soluble solids (TSS) displayed in degrees brix. On the other hand, the X-axis represents the maturity level of each Montaji Agrihorti lemon. As presented in Figure 4, the Total Soluble Solids

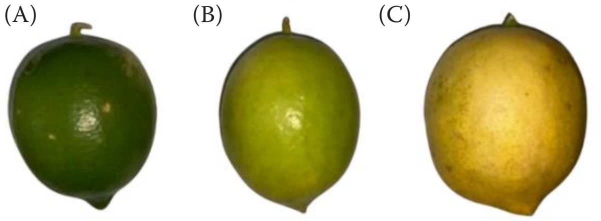

Figure 2. Maturity level of Montaji lemon: (A) unripe, (B) half-ripe, and (C) ripe

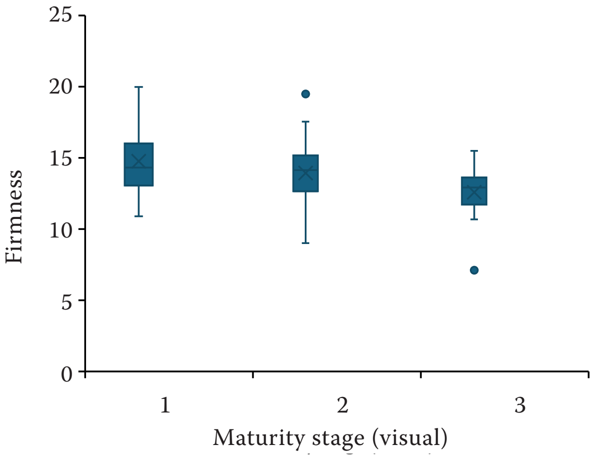

Figure 3. Boxplot of fruit firmness in different maturity level

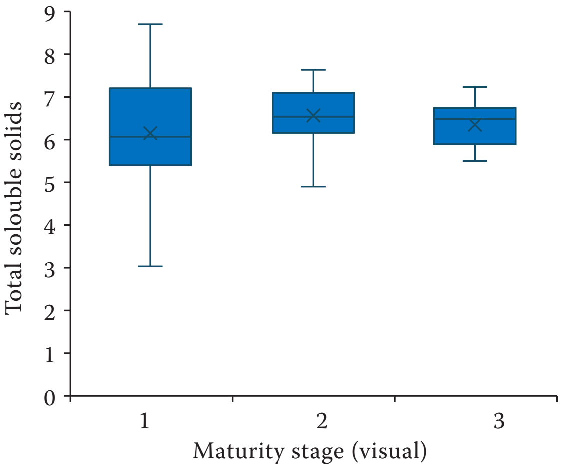

Figure 4. Boxplot of total solouble solids in different maturity level

(Brix) showing increasing trend during ripening. Similar to previous study which asserts that the level of TSS is linear with fruit's age and reaches the highest level at maximum age. The increase in TSS level is attributed to the conversion of starch into sugar, thence, increases the overall TSS level (Ifmalinda et al. 2018).

The relationship between maturity level and acidity. On the other hand, Figure 5 shows the increasing of acidity value during ripening. As referred in the previous study, the lemon acidity increases alongside fruit maturity, resulting in low pH values (Liu et al. 2012). The results of measurement on each Montaji lemon maturity level are presented in Figure 5.

The relationship between maturity level and essential oil yields. The essential oil yield was obtained through the distillation process of Montaji lemon at each level of maturity. For a more detailed description, the measurement results for each maturity level of Montaji lemon are illustrated in Figure 6. Each maturity level acquired a specific essential oil yield of 0.89 for index 1 (unripe) with a total of 2.08 g of oil. At index 2 (half-ripe), the yield was 0.9 with a total of 2.3 g of oil. At index 3 (ripe), the yield was 0.56 with a total of 1.34 g of oil. Selecting harvest time is crucial in determining essential oil yield as well as quality. In practice, in determining essential oil through the distillation process, essential oil data were yielded alongside different contents for each maturity level (Kelen and Tepe 2008).

Pre-processing spectral data. In this study, pre-processing was conducted as an aid to the implemented method to obtain better results. Preprocessing spectral data was conducted by processing a lemon dataset consisting of data obtained from spectrophotometry measurements of Montaji Agrihorti lemons using the smoothing moving average (MA) method. The MA was aimed at reducing data noise, thus producing higher accuracy in the data modeling process. The spectral data from Vis-NIR spectroscopy measurements entail spectral information and noise; hence they require pre-treatment to attain an accurate model (Cen and He 2007). Figure 7 illustrates the moving average as one of the most effective methods for removing spectral noise while preserving the peaks in the spectrum to be subsequently connected to the parameter of Montaji Agrihorti lemons.

Principal component analysis (PCA) and feature selection. The results of spectral data

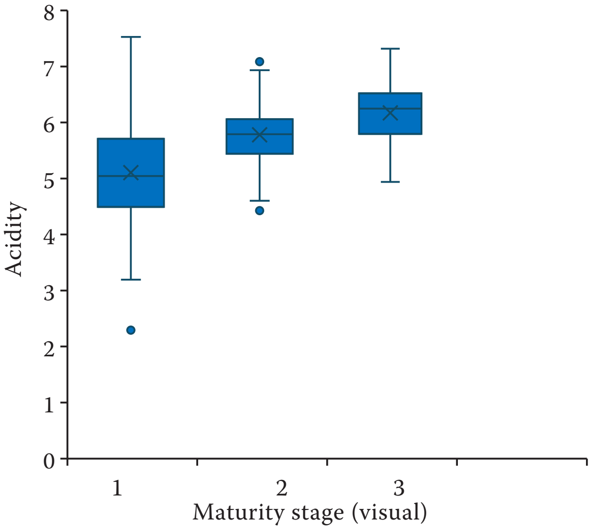

Figure 5. Boxplot of acidity in different maturity level

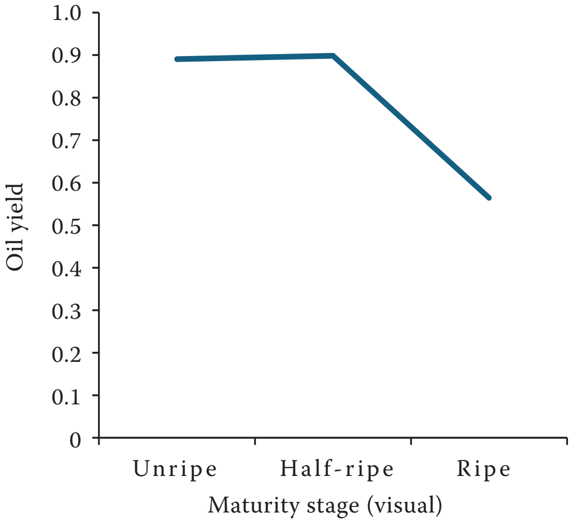

Figure 6. Lineplot of oil yield in different maturity level

https://doi.org/10.17221/25/2024-RAE

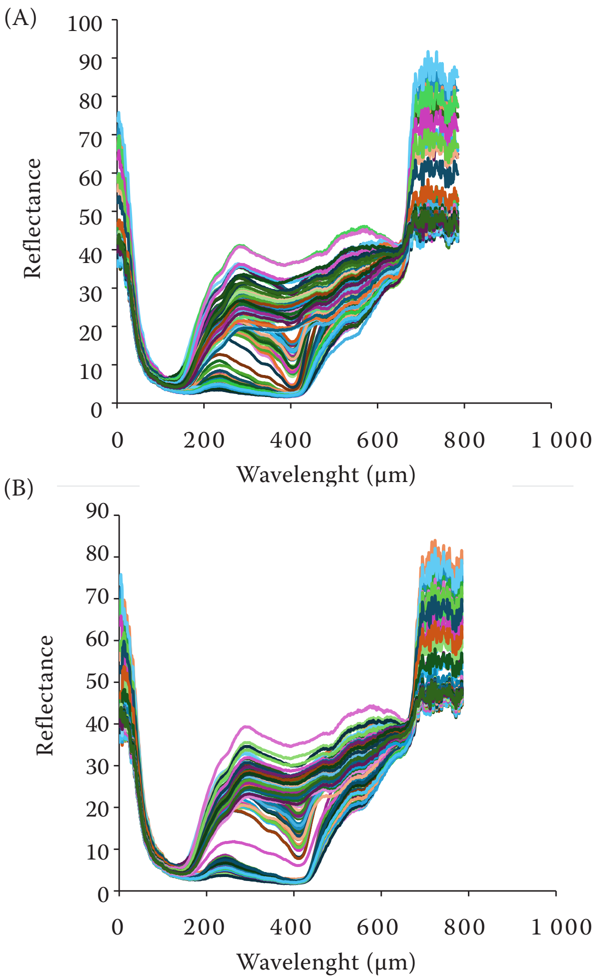

Figure 7. (A) Original spectra and (B) spectra-MA

classification were utilised to attain new dataset prior to conducting feature selection. Feature selection was conducted using principal component analysis (PCA). PCA screeplot was presented in Figure 8. It only needs two components only to reach more than 90.4% of explained variance, with the highest PCA scores indicating PC1 (61.2%) and PC2 (29.3%). The determination of the number of PC can be based on the cumulative proportion of the original data variability by n-principal components, with a minimum of 80% (Hasbullah and Ismail 2022).

Feature selection is the process of selecting evaluation criteria to attain an optimal set of features. This process involves discovering correlated features. Selection is conducted on features that do not provide useful information and are irrelevant, aka noise, and reduce classification performance. Feature selection is performed using PC (principal component). Feature selection is one of the essential steps commonly implemented in data

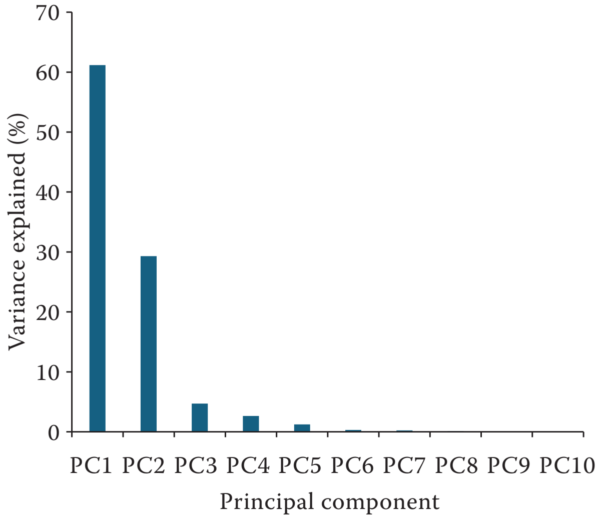

Figure 8. Principal component analysis screeplot

pre-processing. This process is aimed at reducing the number of inappropriate features or data, so as to significantly improve accuracy compared to classification that excludes feature selection (Nguyen et al. 2020). For a clearer visualisation, the selected features are illustrated in Figure 9.

Feature selection was performed using initial spectral data of 785 features, which was subsequently reduced to 36 features. The selected wavelength regions include *A* = 471.1–472.6 nm; *B* = 500.4–545.2 nm; *C* = 636.7–638.1 nm; *D* = 721.7–798.5; *E* = 850.4–868.7 nm. The selected features were mainly located in the regions of 471.1–638.1 nm. As referred

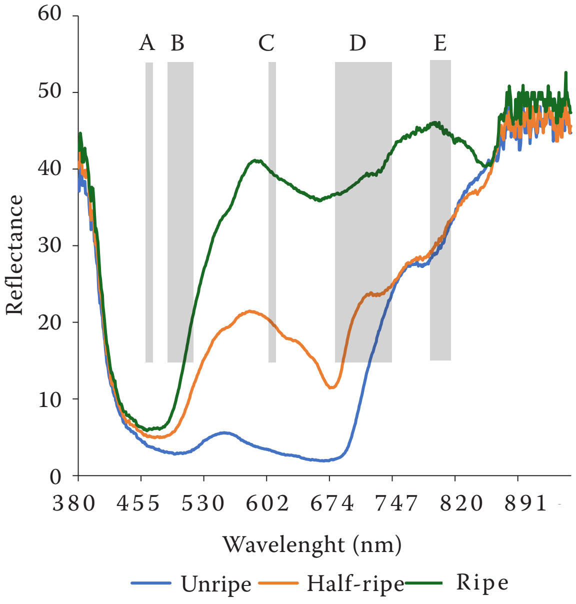

Figure 9. Selected features visualisation

Letters represent selected features region

https://doi.org/10.17221/25/2024-RAE

to in the previous study (Luo et al. 2022), regardless of the three index classifications, the differences between different indexes mainly resulted from changes in colour and texture. Additionally, the selection process was influenced by changes in pigment content during maturation, such as chlorophyll, carotenoids, and anthocyanins in the fruit peels, and strengthened by the NIR spectrum that transferred information on scattering and absorption.

Classification results using machine learning model. The classification results of Montaji Agrihorti lemons from a total sample of 120 fruits from unripe to ripe were conducted using machine learning method. The classification process involved training and testing. The study implemented 70% of the training data and 30% of testing data. The classification process employed various models, including SVM, K-Nearest Neighbors Algorithm (KNN), and Random Forest. In this case, different accuracies for each model with a different algorithm could be observed for each physicochemical parameter of Montaji Agrihorti lemon, as presented in Table 1. The calculation of accuracy was attained from confusion matrix, a commonly applied method to calculate data accuracy.

The best model for maturity parameter based on visual and acidity, with all features or selected features, produced the highest accuracy using random forest. This is due to the fact that the model consistently produced the highest accuracy values compared to the other two classification models. In addition, the model produced low misclassified values compared to other models. The visual parameter, before and after feature selection, declined to 36 features, with training accuracy of 100%, and testing accuracy of 97%. In the visual parameter, the

Table 1. Accuracy results from machine learning model

Parameter	Feature	Model	Accuracy (%)	
			training	testing
Visual	All features-MA	SVM	93	94
		KNN	98	97
		Random forest	100	97
	36-PCA-MA	SVM	92	86
		KNN	98	97
		Random forest	100	97
TSS	All features-MA	SVM	82	72
		KNN	85	72
		Random forest	100	67
	36-PCA-MA	SVM	82	72
		KNN	86	72
		Random forest	100	67
Acidity	All features-MA	SVM	74	86
		KNN	81	86
		Random forest	100	81
	36-PCA-MA	SVM	74	81
		KNN	76	81
		Random forest	100	83
Firmness	All features-MA	SVM	82	78
		KNN	87	75
		Random forest	100	72
	36-PCA-MA	SVM	82	78
		KNN	85	75
		Random forest	100	69

SVM – support vector machine; KNN – K-nearest neighbors algorithm; MA – moving averages; PCA – principal component analysis; TSS – total soluble solids attained accuracy demonstrated a significant influence on the classification of fruit maturity. For the acidity parameter, the training accuracy was 100%, while testing accuracy was 81% for all features.

On the other hand, with feature selection, the training accuracy was 100% and the testing accuracy was 83%, indicating that the testing accuracy was higher compared to the classification that excluded feature selection. As reported in a previous study (Kasimati et al. 2022), the advantages of random forests involve the ability to improve accuracy when encountering missing data and the efficiency of storing large datasets. Additionally, random forest incorporates a feature selection process that is capable of selecting the best feature, thus enhancing classification performance. Different from the visual and acidity parameters, the brix parameter, and firmness with all features and selected data, the highest accuracy was exhibited by the SVM model.

As for the brix parameter, before and after feature selection, there were 36 features, with training accuracy of 82% and testing accuracy of 72%. For the firmness parameter, training accuracy was 82%, and testing accuracy was 78% for all features, in which the testing accuracy was higher compared to selection that excluded feature selection. The results of the classification accuracy generated through the confusion matrix indicated how well the classification was performed. The classification model performed well on classifying the level of maturity of Montaji Agrihorti lemons; thus, both methods – SVM and random forest could demonstrate the ability in distinguishing classes. One of the considerable values is the accuracy value, denoted as C (Permana et al. 2020).

CONCLUSION

Findings of the study suggest that visually observed lemon maturity is correlated to the maturity parameter measured destructively. For firmness, the lemons are linear with the increase in maturity index. The brix index is linear to the fruit's age and reaches its highest level at the maximum age. On the other hand, the acidity index experiences an increase in acidity as the lemons mature. Correspondingly, a prediction model based on Vis-NIR spectroscopy data was successfully created. The modeling results with the best accuracy are exhibited through Random Forest for visual and acidity parameters. For the visual parameter, the testing accuracy is 97% for all features and 36 selected features. As for acidity, testing accuracy is 81% for all features and 83% for 36 features. In addition to random forest, the SVM model produces the best accuracy for brix and firmness parameters. For the Brix parameter, testing accuracy for all features and 36 features is 72%. For firmness, testing accuracy is 78% for all features, or 36 features. Accuracy results for visual, acidity, and firmness parameters signify how well the models distinguish classes, in which the accuracy value is determined by how well the model distinguishes the classes.

REFERENCES

Bhargava A., Bansal A. (2021): Fruits and vegetables quality evaluation using computer vision: A review. Journal of King Saud University – Computer and Information Sciences, 33: 243–257.

Cen H., He Y (2007): Theory and application of near infrared reflectance spectroscopy in determination of food quality. Trends in Food Science & Technology, 18: 72–83.

Dogan N., Dogan I. (2010): Determination of the number of bins/classes used in histograms and frequency tables: A short bibliography. Journal of Statistical Research, 7: 77–85.

Hadiwijaya Y., Kusumiyati K., Munawar A.A. (2020): Application of visible-near infrared spectroscopy technology for rapid and simultaneous prediction of water content in golden melon (*Cucumis melo* L.) fruit. Agroteknika, 3: 67–74.

Hasbullah R., Ismail E.R. (2022): Shelf-life prediction of *Citrus limon* using a multivariate accelerated shelf-life testing (maslt) approach. Journal of Horticultural Research, 30: 51–60.

Ifmalinda Fahmy K., Fitria E. (2018): Prediction of siam gunung omeh citrus fruit (*Citrus nobilis* var Microcarpa) maturity using image processing. Jurnal Keteknikan Pertanian, 6: 335–342.

Kasimati A., Espejo-García B., Darra N., Fountas S. (2022): Predicting grape sugar content under quality attributes using normalized difference vegetation index data and automated machine learning. Sensors, 22: 3249.

Kelen M., Tepe B. (2008): Chemical composition, antioxidant and antimicrobial properties of the essential oils of three *Salvia* species from Turkish flora. Bioresource Technology, 99: 4096–4104.

Kieling D.D., Prudencio S.H. (2018): Blends of lemongrass derivatives and lime for the preparation of mixed beverages: Aantioxidant, physicochemical, and sensory properties. Journal of the Science of Food and Agriculture, 99: 1302–1310.

https://doi.org/10.17221/25/2024-RAE

Lim J., Mo C., Kim G., Kang S., Lee K., Kim M. S., Moon J. (2014): Non-destructive and rapid prediction of moisture content in red pepper (*Capsicum annuum* L.) powder using near-infrared spectroscopy and a partial least squares regression model. Journal of Biosystems Engineering, 39: 184–193.

Liu Y., Heying E., Tanumihardjo S.A. (2012): History, global distribution, and nutritional importance of citrus fruits. Comprehensive Reviews in Food Science and Food Satefy, 11: 530–545.

Luo W., Fan G., Tian P., Dong W., Zhang H., Zhan B. (2022): Spectrum classification of citrus tissues infected by fungi and multispectral image identification of early rotten oranges. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 279: 121412.

Nguyen C.-N., Phan Q.-T., Tran N.-T., Fukuzawa M., Nguyen P.-L., Nguyen C.-N. (2020): Precise sweetness grading of mangoes (*Mangifera indica* L.) based on random forest technique with low-cost multispectral sensors. IEEE Access, 8: 212371–212382.

Nurlatipah S., Halim Y., Gunawan Sebayang R. (2023): Effectiveness of applying anti-aging cream with lemon peel extract (*Citrus limon*) in inhibiting the skin aging process of wistar rats (*Rattus norvegicus*) exposed to ultra violet-B rays. International Journal of Health and Pharmaceutical (IJHP), 3: 973–979.

Permana B.R.S., Panudju A.T. (2020): Comparison and performance analysis of svm and pso-svm algorithms (Case Study classification of senior high school). International Journal of Research and Innovation in Applied Science (IJRIAS). 5: 81–88.

Rafique S., Syeda Mona H., Shahzad Sharif M., Syed Khurram H., Nageena S., Sumaira P., Maryam M., Muhammad F. (2020): Biological attributes of lemon: A review. Journal of Addication Medicine and Therapeutic Science, 6: 30–34.

Samaradiwakara S.D., Champa W.A.H., Eeswara J.P. (2019): Harvest maturity affects postharvest quality of lime fruits (*Citrus aurantifolia* Swingle). Tropical Agricultural Research, 30: 125.

Zhao W., Adolph A.L., Puyau M.R., Vohra F.A., Butte N.F., Zakeri I.F. (2013): Support vector machines classifiers of physical activities in preschoolers. Physiological Reports, 1: e00006.

Received: March 4, 2024
Accepted: October 18, 2024
Published online: December 31, 2024