Efek Spectral

by Ali Khumaidi

Submission date: 26-Nov-2021 04:16PM (UTC+0700)
Submission ID: 1713035219
File name: 16001-35612-3-RV.docx (730.94K)
Word count: 5406
Character count: 29954

EFFECTS OF OVERSAMPLING SMOTE AND SPECTRAL TRANSFORMATIONS IN THE CLASSIFICATION OF MANGO CULTIVARS USING NEAR-INFRARED SPECTROSCOPY

Ali Khumaidi [a,*], Ridwan Raafi'udin [b]

[a] *Faculty of Engineering, Universitas Krisnadwipayana, Jakarta, 17411, Indonesia*
[b] *Faculty of Computer Science, Universitas Pembangunan Nasional Veteran Jakarta, City, 12450, Indonesia*
[*] *Corresponding author: alikhumaidi@unkris.ac.id*

***Abstract*— Near-Infrared spectroscopy (NIR) is a non-destructive analytical technique that can provide chemical and structural information on samples in a speedy and accurate time. NIR has a wavelength of 750-2500 nm. However, the absorbance bands of the NIR spectrum are often broad, non-specific, and overlapping. NIR spectrum analysis requires a multivariate method which is very subjective to noise arising from instrumentation. There is no standard protocol in modelling for classification and prediction using NIR spectra. Several models have been developed with and without pre-processing techniques. The SMOTE technique can improve the model so that it can accurately predict all class responses. This research contributes to creating a multiclass classification model for grouping mango cultivars by finding the best pre-processing technique and using SMOTE oversampling. The results of the four test scenarios on the model's performance built using the Support Vector Machine (SVM) that the best model is obtained using spectral transformations with LSNV and CLIP operations with 100% accuracy, precision and recall values. The Decision Tree (DT) has the performance results in 100% model was obtained by using spectral transformation with LSNV, CLIP and SAVGOL operations with parameters {'deriv_order': 0,1, 2, 'filter_win': 11, 13, 'poly_order': 3}. Using of SMOTE has better accuracy than without preprocessing with an accuracy of 92% on SVM and 94% on DT. While combination of SMOTE and Spectral Transformation gives classification results for SVM and DT with the same accuracy of 96%, better than using SMOTE only.**

***Keywords*— Classification; Cultivar Mango; Near-Infrared; Spectral Transformation; Oversampling SMOTE.**

Manuscript received 15 Oct. 2020; revised 29 Jan. 2021; accepted 2 Feb. 2021. Date of publication 17 Feb. 2021.

I. INTRODUCTION

Technological advances have encouraged innovation in technology development for determining the characteristics and quality of fruit non-destructively [1]. This technology can quantify fruit quality based on the classification of external and internal factors that the human senses cannot detect. The current quality determination method is mostly carried out destructively, which requires time, effort, cost, and there is a bias factor due to human subjectivity [2]. So the method of measuring quality and destructive detection is not suitable to be applied in industry. The non-destructive quality measurement method is more effective based on the correlation between the physical properties of the fruit associated with the fruit quality factor. The use of non-destructive equipment produces more consistent results than human labour, thereby minimizing the chance of errors in determining fruit quality [3].

The application of Near Infrared Spectroscopy (NIR) by utilizing infrared rays is not new, and NIR spectroscopy was developed in 1950 in the industrial field, which focused on analyzing the chemical content of materials [4]. NIR spectroscopy has begun to be widely used to analyze moisture, protein and fat content in agricultural and food products. NIR spectroscopy is a non-destructive analytical technique capable of providing chemical and structural information on certain samples in a swift time (less than 1 minute). NIR has a wavelength of 750-2500 nm, the target sample is illuminated with light, and the reflected light or backscatter is measured with a spectrometer. Horticultural products can also take advantage of this NIR method in grading, sorting, internal quality, and determining harvest time [5]. Determination of the quality of horticultural products can be done non-destructively by using NIR spectroscopy that has been applied to watermelon [6] and melon [7], using spectral for detection Lycopene Content in Tomato [8], detection of mango quality [9], [10].

There is no standard protocol in modelling for classification and prediction using the NIR spectrum. Several models have been developed with and without pre-processing techniques. Support Vector Machine (SVM) is the most widely used algorithm in prediction models, both classification and regression for fruit quality detection and has fairly good accuracy. Detection of black tea quality using standard normal variate (SNV) spectral transformation and Savitzky Golay combination with first derivative and SVM algorithm [11]. Pear hardness measurement using SNV spectral transformation and first-derivative with SVM [12].

Spectral transformation technique can improve model performance [13]. These techniques include Smoothing, Scatter Correction, Trimming, Clipping, Resampling and Derivatives. The order of pre-processing operations applied can affect the performance of the model [14]. Smoothing aims to smooth the spectral and help remove noise. Scatter correction aims to counteract the effects of particle size. Trimming allows the extraction of continuous and non-continuous wavelength regions from full spectral data. Clipping aims to remove or replace data points with values that exceed a user-defined threshold. Resampling processes a new spectral resolution using the Fourier method, which can combine the obtained spectral with several devices having different spectral resolutions.

A Balance class is a condition of unbalanced distribution between classes in a dataset, where one class has a very large amount of data (majority class) compared to the other class (minority class) [15]. The large difference in the amount of data between classes can result in the classification model often not being able to predict the minority class correctly so that a lot of test data that should be in the minority class is predicted wrongly by the classification model [16]. To overcome the imbalanced class problem, one of the methods used is sampling. The sampling method modifies the distribution of data between the majority and minority classes in the training dataset to balance the amount of data for each class [17]. One of the sampling methods that is often used is the Synthetic Minority Oversampling Technique (SMOTE).

The aims of this study are (1) to investigate the effect of the performance of the spectral transformation method and the most optimal operation on the dataset; (2) to find out the effect of data balance on the model; (3) to explore the optimal machine learning classification model based on the application of spectral transformation and data balance.

II. MATERIALS AND METHODS

The research stages include dataset preparation, pre-processing or spectral transformation to obtain the most optimal data to support the model, SMOTE to deal with class imbalances, splitting the data into training data and testing data, developing models and evaluating the model to find out the best model. The modelling will compare the support vector machine (SVM) and the decision tree algorithm. To determine the performance of the classification model by evaluating the model with four scenarios are (1) without any treatment; (2) apply SMOTE; (3) apply spectral transformation; (4) apply SMOTE and spectral transformation. The relationship between stages can be seen in Figure 1.

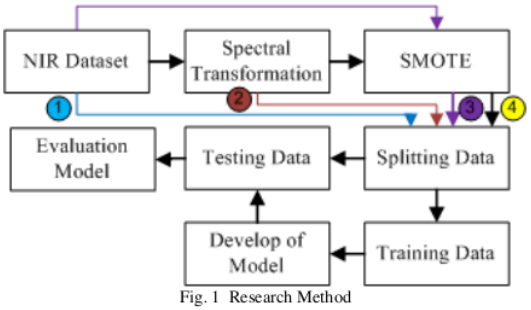


Fig. 1 Research Method

A. Dataset

The dataset used is derived from research results [18]. A total of 186 whole mango samples from 4 different cultivars (Cengkir, Kweni, Kent and Palmer) were taken using near-infrared spectral obtained in the form of absorbance with wavelengths from 1000 to 2500 nm. The number of samples of cultivar Cengkir 18, Kweni 29, Kent 85 and Palmer 54. This dataset is accessible https://data.mendeley.com/datasets/b9d6s7hr33/1.

B. Pre-processing

The purpose of data pre-processing is to reduce the physical properties in the spectrum so as to reduce the variability caused by light scattering, nonlinearity and improve the model to be used [19]. The spectral transformations used are resampling, clipping, smoothing, derivative and scatter correction in smoothing using Savitsky Golay filtering with parameters: filter windows (7, 11, 13), order of the polynomial = 3, and order of the derivative (0, 1, 2). Scatter correction uses several operations, are multiplicative scatter correction (MSC), standard normal variate (SNV), robust normal variate (RNV), normalization, baseline, detrend, localized version of SNV (LSNV) and The extended version of MSC (EMSC). The use of the spectral transformation method in more detail can be seen in Table 1.

The MSC method is to draw spectral sample points to approach the reference spectrum by utilizing the results of estimating simple linear regression parameters and can eliminate the variation between spectra by correcting the scatter position of each intensity value of each replication to the scattering position of the average intensity of the entire replication [20]. The SNV method removes the scattering effect from the spectrum by centering and adjusting the scale of each spectrum [21]. The RNV method is more suitable for data with a lot of noise by using the concept of correction based on the median value and the interval between quartiles[22]. The spectral normalization method uses a certain range of values and usually applies to Euclidean. The baseline method principally uses the average of the central values of the spectral. The concept of the LSNV method is similar to that of SNV with the principle of division operation on the spectral window. The EMSC method is in principle almost the same as the MSC, but in EMCS, it takes into account linear and quadratic corrections.

All methods and operations of this spectral transformation will be compared to the model to obtain the most optimal accuracy—the use of spectral transformation techniques as one of the factors in improving the performance of the model.

TABLE I
SPECTRAL TRANSFORM METHODS

Method	Operation	Parameter	Value
Resampling	RESAMPLE	Rasio	0.8
Clipping	CLIP	Threshold	1e4
		substitute	None
Smoothing	SAVGOL	filter_win	7, 11, 13
		poly_order	3
		deriv_order	0, 1, 2
Scatter Correction	MSC		
	SNV		
	RNV	iqr	75-25, 90-10
	NORML		
	BASELINE		
	LSNV		
	EMSC		

C. SMOTE

In this study, it was identified that the dataset used has class imbalance problems so that an over-sampling method is needed to overcome the imbalanced class problem. The method that can be used is SMOTE. SMOTE is an over-sampling method in which the data in the minority class is reproduced using synthetic data derived from data replication in the minority class. Over-sampling in SMOTE takes an instance of the minority class and then looks for the k-nearest neighbour of each instance, then generates a synthetic instance instead of replicating the minority class instance; therefore, it can avoid the problem of excessive overfitting [23]. The algorithm that works on the first SMOTE will make the difference between the vectors of the features in the minority class and the nearest neighbour values from the minority class and then multiply that value by a random number between 0 to 1. Next, the calculation results are added to the feature vector so that the vector value results are obtained from the new one [24].

The proposed model will be validated by two experimental scenarios that were carried out, namely using the SVM and Decision Tree algorithm approaches and each was used for modelling without considering class imbalance, and secondly, SMOTE oversampling was carried out to increase the number of datasets in order to achieve a balanced dataset.

D. Modelling

The modelling develops in this research is classification. The developing model will group based on NIR spectral in 4 classes of mango cultivar. Classification is a multivariate technique for separating different sets of objects and allocating new objects into predefined groups. A good classification method will result in less misclassification. It is necessary to use the right method to perform the classification accurately, it. Support Vector Machine (SVM) is one method that can perform classification. SVM is a technique for finding hyperplanes that can separate two sets of data from two different classes [25]. SVM has advantages including in determining the distance using a support vector so that the computational process becomes fast. The learning process in SVM aims to obtain a hypothesis in the form of the best dividing field that minimizes not only empirical risk, namely the average error in the training data, but also provides good generalization. Generalization is the ability of a hypothesis to be able to classify data that is not contained in the training data correctly. The principle of SVM is actually a linear classifier and then redeveloped so that it can work on non-linear problems using the kernel trick method, which is looking for a hyperplane by transforming the dataset into a vector space with larger dimensions (feature space) using a kernel function which will then be classified and performed on the feature space. Determination of the kernel function used will greatly affect the classification results [26].

The decision tree is one of the most popular classification methods because it is easy for humans to interpret. The decision tree is a predictive model using a tree structure or hierarchical structure [27]. The concept of a decision tree is to convert data into a decision tree and decision rules. A Decision Tree is used to study the classification and prediction of patterns from data and describe the relationship of the attribute variable x and the target variable y in the form of a tree. The decision tree resembles a flowchart where each internal node (a node that is not a leaf or the outermost node) is a test of attribute variables; each branch is the result of the test, while the outermost node, namely the leaf, is the. The main benefit of using a decision tree is its ability to break down complex decision-making processes into simpler ones so that decision-makers will better interpret solutions to problems [28].

Decision trees are also useful for exploring data, finding hidden relationships between a number of potential input variables and a target variable. Decision trees combine data exploration and modelling, so they are great as a first step in the modelling process even when used as the final model of some other technique. Another advantage of this method is that it can eliminate unnecessary calculations or data. This is because the existing samples are usually only tested based on certain criteria or classes [29].

E. Model evaluation

A confusion matrix will be used in measuring the performance of the classification model using a. The confusion matrix, also known as the error matrix, provides information on the comparison of the classification results performed by the model with the actual classification results [30]. There are four terms representing the results of the classification process in the confusion matrix, namely True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). TP is positive data that is predicted to be correct, TN is negative data that is predicted to be correct, FP is negative data but is predicted to be positive data and FN is positive data but is predicted to be negative data.

The confusion matrix can calculate various performance metrics to measure the performance of the model that has been made, some of which are often used, namely accuracy, precision, and recall. Accuracy describes how accurately the model can classify correctly. Precision describes the level of accuracy between the requested data and the prediction results provided by the model. Recall describes the success of the model in retrieving information.

III. RESULT AND DISCUSSION

The NIR spectral dataset of 186 samples with a wavelength of 1000 to 2500 nm can be seen in Figure 4a. There are several absorption peaks that can be found from the original transmittance spectrum. The mango dataset consists of 4 cultivars so that the multiclass classification method was used in this study. The following are the results of measuring the

model performance from the four scenarios with the distribution of training data and testing data of 70 and 30.

A. Modelling without pre-processing

The results of the classification of NIR spectrum data processing using SVM without pre-processing have resulted in a fairly good accuracy of 90%, class classification errors of 6 Kent cultivars which are actually Palmer cultivars. The results of the classification using DT without pre-processing resulted in higher accuracy than SVM, which was 94%, with an incorrect guess of 4 Cengkir cultivars, which should have been Palmer cultivar. The results of the confusion matrix in more detail can be seen in Table 2.

TABLE II
CONFUSION MATRIX RESULTS WITHOUT PRE-PROCESSING AND USING SPECTRAL TRANSFORMATION

Algorithm	Class	Prediction				Without Pre-Processing			Prediction				Using Spectral Transformation		
		C	K	Kw	P	Accuracy	Precision	Recall	C	K	Kw	P	Accuracy	Precision	Recall
SVM	Cengkir (C)	6	0	0	0	0,90	0,96	0,92	5	0	0	0	1	1	1
	Kent (K)	0	28	0	0				0	26	0	0			
	Kweni (Kw)	0	0	10	0				0	0	13	0			
	Palmer (P)	0	6	0	12				0	0	0	18			
DT	Cengkir (C)	6	0	0	0	0,94	0,9	0,94	5	0	0	0	1	1	1
	Kent (K)	0	28	0	0				0	26	0	0			
	Kweni (Kw)	0	0	10	0				0	0	13	0			
	Palmer (P)	4	0	0	14				0	0	0	18			

B. Modelling with oversampling SMOTE

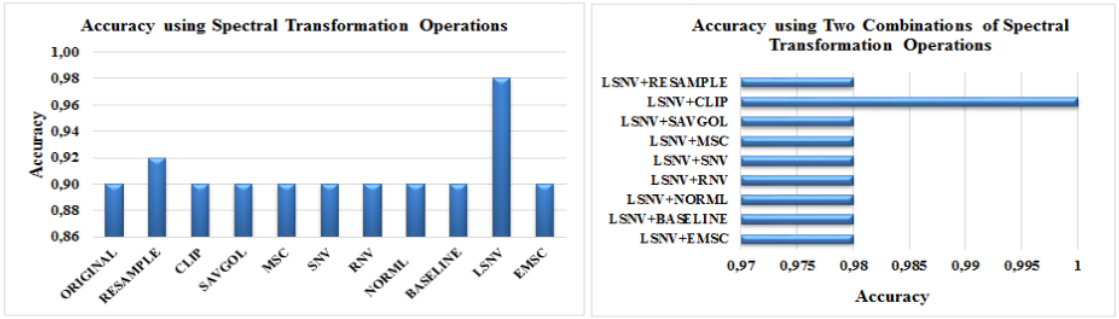
With the application of SMOTE oversampling, the amount of data between classes is balanced with the number of 85 mangoes in each class. Cengkir cultivars, which originally had 18 data, Kweni had 29 data, and Palmer originally had 54 data, were equated with the total data of Kent, which was 85 data. The results of the classification of NIR spectrum data processing using SVM with the application of SMOTE oversampling improved the classification accuracy of 92% compared to those without pre-processing, the class grouping error of 1 Palmer cultivar, which was actually a Cengkir cultivar, 1 Cengkir cultivar which was actually a Kent cultivar and the supposed Palmer cultivar was predicted to be 2 cultivars Cup and 5 Kent.

The results of the classification using DT with the application of SMOTE oversampling improved the classification accuracy of 94% compared to those without pre-processing, class classification errors were 1 Palmer cultivar and 1 Kent cultivar which was actually Cengkir cultivar, 3 Palmer cultivar which was actually Kent cultivar and 1 Cengkir cultivar which was supposed to be Palmer. The results of the confusion matrix in more detail can be seen in Table 3.

C. Modelling with spectral transformation

The results of the classification of NIR spectrum data processing using SVM with the application of spectral transformation methods, namely Smoothing, Scatter Correction, Clipping, Resampling and Derivatives and their combinations, operations and parameters of these methods, can be seen in Table 1. Using the Nippy library with Python, the results obtained classification with an accuracy value of 100%, where the use of the Clipping and Scatter Correction method with LSNV operation provides the most optimal results for SVM. The results of the DT classification with spectral transformation using the Clipping and Scatter Correction method with LSNV operation also produce 100% accuracy without class prediction errors using the Clipping, Scatter Correction and Smoothing methods. The results of the confusion matrix in more detail can be seen in Table 2.

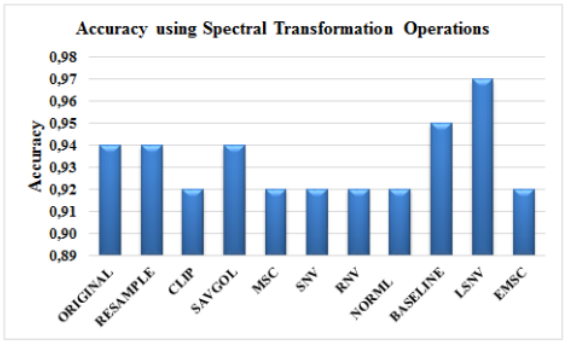
The classification accuracy calculation process using SVM and DT with a combination of 5 spectral transformation methods gave varying results. Some even had worse accuracy with accuracy values without pre-processing. Therefore, it is very important to adapt spectral transformation methods and operations to the data. Experiments were carried out on all spectral transformation operations to obtain the most optimal accuracy results in the classification using SVM. First, testing is carried out on each operation and its parameters. The results of the spectral transformation operation with the best accuracy are then combined with other operations. The results of 2 combinations of spectral transformation operations are then combined with other operations. And so on until the most optimal accuracy value is obtained.



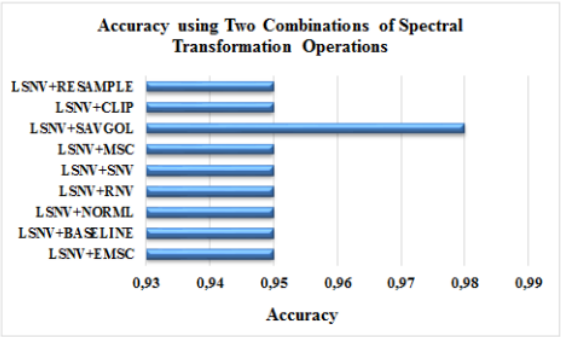
(a) (b)

Fig. 2 Comparison of Combination of Spectral Transformation toward Accuracy on SVM

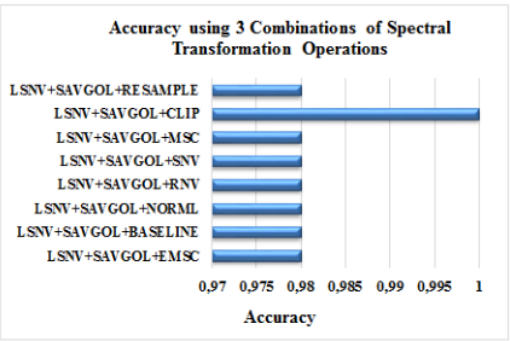
Figure 2a describes the results of applying one spectral transformation operation on SVM. Spectral transformation operations that produce higher accuracy values than those without pre-processing are RESAMPLE with 92% accuracy and LSNV with 98% accuracy. For other operations, the accuracy value is 90% the same as without pre-processing. The LSNV operation as the spectral transformation operation with the highest value is then combined with other operations. Accuracy results from 2 combinations of spectral transformation operations on SVM have reached 100% accuracy by using LSNV and CLIP operations with threshold=10000. While the other combinations with a maximum accuracy value of 98%. By having obtained the most optimal spectral transformation operation, the next combination will achieve optimal again, at least by using two combinations of spectral transformation operations, namely LSNV and CLIP. The accuracy results of 2 combinations of spectral transformation operations on SVM can be seen in Figure 2b.



(a)



(b)



(c)

Fig. 3 Comparison of Combination of Spectral Transformation toward Accuracy on DT

Figure 3a describes the results of applying one spectral transformation operation to DT. Spectral transformation operations that produce higher accuracy values than those without pre-processing are BASELINE with 95% accuracy and LSNV with 97% accuracy. For other spectral transformation operations, the accuracy value is lower than the accuracy value without pre-processing, which is 92%. The LSNV operation as the spectral transformation operation with the highest value is then combined with other operations. Accuracy results from 2 combinations of spectral transformation operations on DT that have increased, namely LSNV and SAVGOL operations with parameters {'deriv_order': 2, 'filter_win': 11, 'poly_order': 3} with an accuracy value of 98%. While the combination of LSNV with other spectral transformation operations, the maximum classification accuracy value is 95%. Accuracy results with two combinations of spectral transformation operations on DT can be seen in Figure 3b. Next is to perform three combinations of spectral transformation operations using LSNV and SAVGOL as operations with the best accuracy on two combinations of spectral transformation operations combined with other spectral transformation operations. The combination of 3 spectral transformation operations has achieved 100% accuracy by using LSNV, CLIP with threshold=10000 and SAVGOL with parameters {'deriv_order': 2, 'filter_win': 11, 'poly_order': 3}, {'deriv_order' : 1, 'filter_win': 11, 'poly_order': 3}, {'deriv_order': 2, 'filter_win': 13, 'poly_order': 3}, and {'deriv_order': 0, 'filter_win': 13, 'poly_order': 3}. Three other combinations of spectral transformations with a maximum accuracy value of 98%. By having obtained the most optimal spectral transformation operation, the next combination will achieve optimal again, at least by using three combinations of spectral transformation operations, namely LSNV, CLIP and SAVGOL. For the accuracy results of 3 combinations of spectral transformation operations on DT, it can be seen in Figure 3c.

TABLE III
CONFUSION MATRIX RESULTS USING OVERSAMPLING SMOTE AND SPECTRAL TRANSFORMATION

Algorithm	Class	Prediction				Using Oversampling SMOTE			Prediction				Using Oversampling SMOTE and Spectral Transformation		
		C	K	Kw	P	Accuracy	Precision	Recall	C	K	Kw	P	Accuracy	Precision	Recall
SVM	Cengkir (C)	27	0	0	1	0,92	0,92	0,92	25	3	0	0	0,96	0,97	0,96
	Kent (K)	1	27	0	0				0	28	0	0			
	Kweni (Kw)	0	0	29	0				0	0	29	0			
	Palmer (P)	2	5	0	21				0	1	0	27			
DT	Cengkir (C)	26	1	0	1	0,94	0,94	0,94	27	0	0	1	0,96	0,96	0,96
	Kent (K)	0	25	0	3				0	27	0	1			
	Kweni (Kw)	0	0	29	0				0	0	29	0			
	Palmer (P)	1	0	0	22				0	3	0	25			

D. Modelling with oversampling SMOTE and spectral transformation

The results of the classification of NIR spectrum data processing using SVM with the application of SMOTE oversampling and spectral transformation, where there is a balance of the amount of data in the class and the use of Clipping and Scatter Correction methods with LSNV operations, produces a classification accuracy value of 96%. The class classification error is 3 Kent cultivar which is actually Cengkir cultivar, and 1 Kent cultivar, which is actually Palmer cultivar. The results of the classification using DT also produce a classification accuracy value of 96%. Class classification errors are 1 Palmer cultivar which is actually Cengkir cultivar, 1 Palmer cultivar, which is actually Kent cultivar and 3 Kent cultivar, which should be Palmer. The results of the confusion matrix in more detail can be seen in Table 3. The performance of classification accuracy with SVM and DT has the same value, and both are more or less better than using spectral transformation alone. However, the accuracy value is higher than without pre-processing and the use of oversampling SMOTE.

The spectral transformation has a significant influence in increasing the accuracy of the classification model. In the four test scenarios, the classification results with the application of appropriate spectral transformation methods and operations achieved 100% accuracy. The best spectral transformation results with Clipping and Scatter Correction methods with LSNV operation for modelling with SVM can be seen in Figure 4b, and for DT modelling can be seen in Figure 4c. The spectrum peaks are very clearly visible in both images, making it easier for the model to classify.

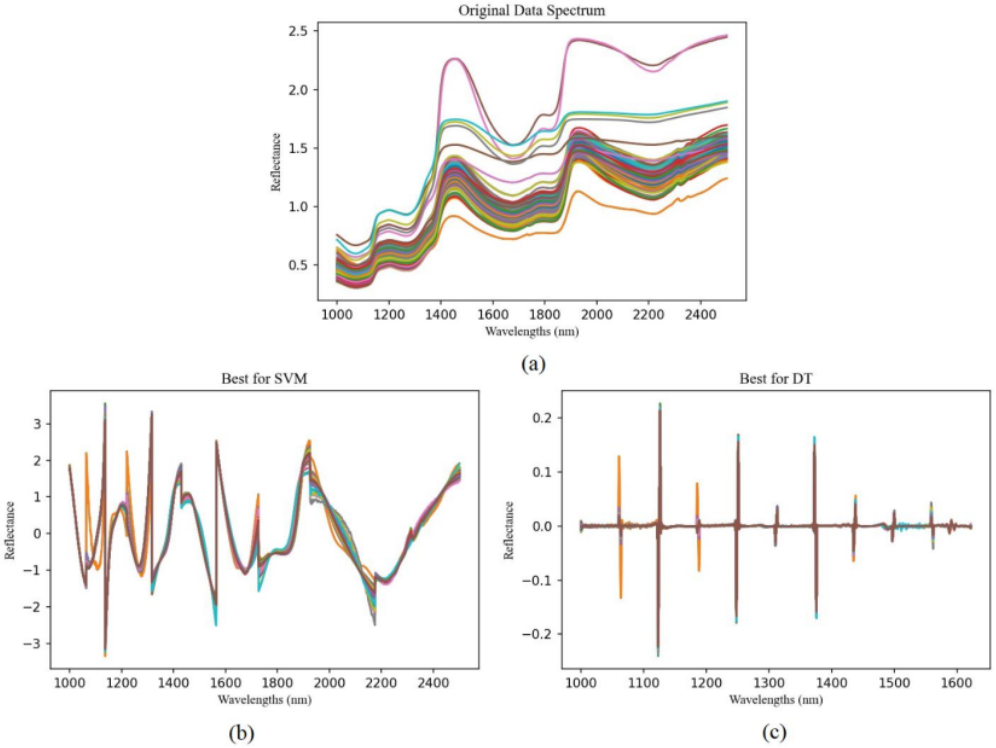


Fig. 4 Comparison of Spectral Curves after using Spectral Transformation

Table 3 is the result of the comparison of 4 scenarios on the SVM and DT algorithms with performance measurements based on accuracy, precision and recall. With the model built by SVM and DT the best performance with the application of spectral transformation then the application of SMOTE oversampling and spectral transformation, then the application of SMOTE oversampling and finally without pre-processing.

IV. CONCLUSION

To improve the accuracy of the classification model of 4 classes of mango cultivars based on NIR spectroscopy, spectral treatment is needed. Treatment with SMOTE oversampling can improve classification accuracy in SVM and DT algorithms. Treatment with spectral transformation using five combinations of methods obtained an optimal classification accuracy of 100% on SVM using clipping and scatter correction methods, namely LSNV, while in DT using clipping, scatter correction and smoothing techniques. When the application of the spectral transformation is optimal with SMOTE oversampling, the classification accuracy is not better. Spectral transformation treatment in classification modelling significantly affects accuracy due to the non-specific, overlapping, broad nature of the NIR spectrum and the presence of noise arising from the instrumentation of the device during data collection. Although the SMOTE oversampling effect causes more data sets and more diverse opportunities for incorrect predictions between classes, it has increased classification accuracy compared to no pre-processing. The SMOTE oversampling technique can be used when there is an imbalance of data classes in the classification.

ACKNOWLEDGEMENT

Thanks to the research results of Mr Munawar et al., who has provided research data, and the head of the faculty of engineering at Krisnadwipayana University, who has supported research and publications.

REFERENCES

[1] P. Osinenko et al., "Application of non-destructive sensors and big data analysis to predict physiological storage disorders and fruit firmness in 'Braeburn' apples," Comput. Electron. Agric., vol. 183, p. 106015, Apr. 2021, doi: 10.1016/j.compag.2021.106015.

[2] M. Arunkumar, A. Rajendran, S. Gunasri, M. Kowsalya, and C. K. Krithika, "Non-destructive fruit maturity detection methodology - A review," Mater. Today Proc., Mar. 2021, doi: 10.1016/j.matpr.2020.12.1094.

[3] B. Nugraha, P. Verboven, S. Janssen, Z. Wang, and B. M. Nicolaï, "Non-destructive porosity mapping of fruit and vegetables using X-ray CT," Postharvest Biol. Technol., vol. 150, pp. 80–88, Apr. 2019, doi: 10.1016/j.postharvbio.2018.12.016.

[4] B. M. Nicolaï et al., "Non-destructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review," Postharvest Biol. Technol., vol. 46, no. 2, pp. 99–118, Nov. 2007, doi: 10.1016/j.postharvbio.2007.06.024.

[5] A. Ibrahim, N. El-Bialee, M. Saad, and E. Romano, "Non-Destructive Quality Inspection of Potato Tubers Using Automated Vision System," Int. J. Adv. Sci. Eng. Inf. Technol., vol. 10, no. 6, p. 2419, Dec. 2020, doi: 10.18517/ijaseit.10.6.13079.

[6] C. Ding, D. Wang, Z. Feng, W. Li, and D. Cui, "Integration of vibration and optical techniques for watermelon firmness assessment," *Comput. Electron. Agric.*, vol. 187, p. 106307, Aug. 2021, doi: 10.1016/j.compag.2021.106307.

[7] J. M. S. Netto, F. A. Honorato, P. M. Azoubel, L. E. Kurozawa, and D. F. Barbin, "Evaluation of melon drying using hyperspectral imaging technique in the near infrared region," LWT, vol. 143, p. 111092, May 2021, doi: 10.1016/j.lwt.2021.111092.

[8] F. D. Anggraeni, N. Khuriyati, M. A. F. Falah, H. Nishina, K. Takayama, and N. Takahashi, "Non-destructive Measurement of Lycopene Content in High Soluble Solids Stored Tomato (Solanum Lycopersicum Mill. cv Rinka 409)," Int. J. Adv. Sci. Eng. Inf. Technol., vol. 10, no. 6, p. 2567, Dec. 2020, doi: 10.18517/ijaseit.10.6.9478.

[9] R. Hayati, A. A. Munawar, and F. Fachruddin, "Enhanced near infrared spectral data to improve prediction accuracy in determining quality parameters of intact mango," Data Br., vol. 30, p. 105571, Jun. 2020, doi: 10.1016/j.dib.2020.105571.

[10] P. Mishra, E. Woltering, and N. El Harchioui, "Improved prediction of 'Kent' mango firmness during ripening by near-infrared spectroscopy supported by interval partial least square regression," Infrared Phys. Technol., vol. 110, p. 103459, Nov. 2020, doi: 10.1016/j.infrared.2020.103459.

[11] G. Ren, Y. Liu, J. Ning, and Z. Zhang, "Assessing black tea quality based on visible–near infrared spectra and kernel-based methods," J. Food Compos. Anal., vol. 98, p. 103810, May 2021, doi: 10.1016/j.jfca.2021.103810.

[12] J. Li, H. Zhang, B. Zhan, Y. Zhang, R. Li, and J. Li, "Non-destructive firmness measurement of the multiple cultivars of pears by Vis-NIR spectroscopy coupled with multivariate calibration analysis and MC-UVE-SPA method," Infrared Phys. Technol., vol. 104, p. 103154, Jan. 2020, doi: 10.1016/j.infrared.2019.103154.

[13] C. Liu, S. X. Yang, X. Li, L. Xu, and L. Deng, "Noise level penalizing robust Gaussian process regression for NIR spectroscopy quantitative analysis," Chemom. Intell. Lab. Syst., vol. 201, p. 104014, Jun. 2020, doi: 10.1016/j.chemolab.2020.104014.

[14] J. Torniainen, I. O. Afara, M. Prakash, J. K. Sarin, L. Stenroth, and J. Töyräs, "Open-source python module for automated pre-processing of near infrared spectroscopic data," Anal. Chim. Acta, vol. 1108, pp. 1–9, Apr. 2020, doi: 10.1016/j.aca.2020.02.030.

[15] Y. Sun, A. K. C. Wong, and M. S. Kamel, "Classification of Imbalanced Data: A Review," Int. J. Pattern Recognit. Artif. Intell., vol. 23, no. 04, pp. 687–719, Jun. 2009, doi: 10.1142/S0218001409007326.

[16] N. S. Sani, M. Abdul Rahman, A. Abu Bakar, S. Sahran, and H. Mohd Sarim, "Machine Learning Approach for Bottom 40 Percent Households (B40) Poverty Classification," Int. J. Adv. Sci. Eng. Inf. Technol., vol. 8, no. 4–2, p. 1698, Sep. 2018, doi: 10.18517/ijaseit.8.4-2.6829.

[17] J. Jang, Y. Kim, K. Choi, and S. Suh, "Sequential targeting: A continual learning approach for data imbalance in text classification," Expert Syst. Appl., vol. 179, p. 115067, Oct. 2021, doi: 10.1016/j.eswa.2021.115067.

[18] A. A. Munawar, Kusumiyati, and D. Wahyuni, "Near infrared spectroscopic data for rapid and simultaneous prediction of quality attributes in intact mango fruits," Data Br., vol. 27, p. 104789, Dec. 2019, doi: 10.1016/j.dib.2019.104789.

[19] Mishra, J. M. Roger, D. N. Rutledge, and E. Woltering, "SPORT pre-processing can improve near-infrared quality prediction models for fresh fruits and agro-materials," Postharvest Biol. Technol., vol. 168, p. 111271, Oct. 2020, doi: 10.1016/j.postharvbio.2020.111271..

[20] D. D. Silalahi, H. Midi, J. Arasan, M. S. Mustafa, and J.-P. Caliman, "Robust generalized multiplicative scatter correction algorithm on pretreatment of near infrared spectral data," *Vib. Spectrosc.*, vol. 97, pp. 55–65, Jul. 2018, doi: 10.1016/j.vibspec.2018.05.002.

[21] B. Lu et al., "Quantitative NIR spectroscopy determination of coco-peat substrate moisture content: Effect of particle size and non-uniformity," Infrared Phys. Technol., vol. 111, p. 103482, Dec. 2020, doi: 10.1016/j.infrared.2020.103482.

[22] Q. Guo, W. Wu, and D. . Massart, "The robust normal variate transform for pattern recognition with near-infrared data," Anal. Chim. Acta, vol. 382, no. 1–2, pp. 87–103, Feb. 1999, doi: 10.1016/S0003-2670(98)00737-5.

[23] T. Pan, J. Zhao, W. Wu, and J. Yang, "Learning imbalanced datasets based on SMOTE and Gaussian distribution," Inf. Sci. (Ny)., vol. 512, pp. 1214–1233, Feb. 2020, doi: 10.1016/j.ins.2019.10.048.

[24] Asniar, N. U. Maulidevi, and K. Surendro, "SMOTE-LOF for noise identification in imbalanced data classification," J. King Saud Univ. - Comput. Inf. Sci., Feb. 2021, doi: 10.1016/j.jksuci.2021.01.014.

[25] M. Sinambela, M. Situmorang, K. Tarigan, S. Humaidi, and M. Sirait, "Waveforms Classification of Northern Sumatera Earthquakes for New Mini Region Stations Using Support Vector Machine," Int. J. Adv. Sci. Eng. Inf. Technol., vol. 11, no. 2, p. 489, Apr. 2021, doi: 10.18517/ijaseit.11.2.12503.

[26] V. Vapnik and R. Izmailov, "Reinforced SVM method and memorization mechanisms," Pattern Recognit., vol. 119, p. 108018, Nov. 2021, doi: 10.1016/j.patcog.2021.108018.
[27] I. Hasanah, E. Purwanti, and P. Widiyanti, "Design and Implementation of an Early Screening Application for Dengue Fever Patients Using Android-Based Decision Tree C4.5 Method," Int. J. Adv. Sci. Eng. Inf. Technol., vol. 10, no. 6, p. 2237, Dec. 2020, doi: 10.18517/ijaseit.10.6.5771.
[28] J. P. Pinder, "Decision Trees," in Introduction to Business Analytics using Simulation, Elsevier, 2017, pp. 47–69.
[29] P. Arumugam and P. Jose, "Efficient Decision Tree Based Data Selection and Support Vector Machine Classification," Mater. Today Proc., vol. 5, no. 1, pp. 1679–1685, 2018, doi: 10.1016/j.matpr.2017.11.263.
[30] J. Xu, Y. Zhang, and D. Miao, "Three-way confusion matrix for classification: A measure driven view," Inf. Sci. (Ny)., vol. 507, pp. 772–794, Jan. 2020, doi: 10.1016/j.ins.2019.06.064.

Efek Spectral

ORIGINALITY REPORT

8%	6%	8%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Anny Kartika Sari, Faisal Malik Widya Prasetya. "Linear SVM for Classifying Breast Cancer Data Encrypted Using Homomorphic Cryptosystem", 2019 5th International Conference on Science and Technology (ICST), 2019 Publication	1%
2	www.ijcams.com Internet Source	1%
3	ejournal.nusamandiri.ac.id Internet Source	1%
4	S. Anggraini, S. K. Wijaya, Daryono. "Earthquake detection and location for Earthquake Early Warning Using Deep Learning", Journal of Physics: Conference Series, 2021 Publication	1%
5	insightsociety.org Internet Source	1%

6	Ade Chandra Nugraha, Suhono Harso Supangkat, IGB Baskara Nugraha, Harno Trimadi et al. "Detection of Railroad Anomalies using Machine Learning Approach", 2021 International Conference on ICT for Smart Society (ICISS), 2021 Publication	1 %
7	thesai.org Internet Source	1 %
8	Nurfaizah, T Hariguna, Y I Romadon. "The accuracy comparison of vector support machine and decision tree methods in sentiment analysis", Journal of Physics: Conference Series, 2019 Publication	1 %
9	P P Ramadhani, S Hadi. "Text classification on the Instagram caption using support vector machine", Journal of Physics: Conference Series, 2021 Publication	1 %
10	journal.uad.ac.id Internet Source	1 %

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		