p-ISSN: 2252-3006
e-ISSN: 2685-2411

Optimization of Preprocessing Spectra and Modeling Using Machine Learning for Prediction of Agricultural Soil Nutrients

Harry Dwiyana Kartika [a1], Nur Hikmah[a2], Ali Khumaidi[a3]

[a]Program Studi Teknik Informatika, Universitas Krisnadwipayana, Indonesia
e-mail: [1] harry_d@unkris.ac.id, [2]nurhikmah@unkris.ac.id, [3]alikhumaidi@unkris.ac.id

Abstrak

Penelitian yang membahas prediksi unsur hara tanah dengan NIR sebagian besar menggunakan algoritma PLS. Kehadiran machine learning (ML) telah menghasilkan metode pembelajaran terotomasi untuk mencari model optimal. Preprocessing dalam pengembangan model prediksi menjadi bagian penting. Preprocessing spektrum bertujuan untuk menghilangkan sumber varians yang tidak informatif. Penelitian mengenai berbagai metode preprocessing terbaik sering ditentukan melalui trial-and-error. Pendekatan preprocessing dengan membandingkan sejumlah operasi preprocessing namun metode ini kurang efisien. Pada penelitian ini mengusulkan penerapan ML untuk menemukan kombinasi operasi preprocessing terbaik secara cepat dan bersamaan. Hasil pengujian preprocessing menggunakan 12 operator menghasilkan 2.112 kombinasi. Penggunaan teknik preprocessing mampu meningkatkan kinerja pada semua algoritma (RF, SVR, PLS, LR, dan MLP). Pengujian unsur tanah K memiliki error terendah pada algoritma LR, pengujian unsur tanah Mg, Ca, P, dan pH menggunakan algoritma MLP memiliki kinerja terbaik dan pada pengujian unsur tanah N kinerja terbaik pada algoritma RF.

Kata kunci: *prediksi, unsur hara tanah, NIR, preprocessing, machine learning*

Abstract

Research that addresses soil nutrient prediction with NIR mostly uses PLS algorithms. Advent of machine learning (ML) has resulted in automated learning methods to find optimal model. Preprocessing in the development of prediction models is an important part. Spectrum preprocessing aims to eliminate uninformative sources of variance. Research on the best preprocessing methods is often determined through trial-and-error. Preprocessing approach compares a number of preprocessing operations but this method is less efficient. This research proposes the application of ML to find the best combination of preprocessing operations quickly and simultaneously. Preprocessing test results using 12 operators resulted in 2,112 combinations. Use of preprocessing can improve performance of all algorithms (RF, SVR, PLS, LR, and MLP). K soil element testing has lowest error in LR, Mg, Ca, P, and pH soil element testing using MLP has the best performance and in N soil element testing the best performance in RF.

Keywords : *prediction, soil nutrients, NIR, preprocessing, machine learning*

1. Introduction

In precision agriculture soil fertility is an important factor affecting crop growth. Soil fertility should always be monitored in real time by determining its properties such as micro and macro nutrients. Near infrared (NIR) spectroscopy has rapidly developed into a fast and effective analytical method for various fields [1]. NIR spectroscopy is widely proposed as an alternative method for determining nutrients and soil quality properties [2]. Nitrogen (N), Phosphorus (P), Potassium (K), Magnesium (Mg), Calcium (Ca) and soil pH are macro-nutrients needed to support plant growth.

Soil fertility properties and nutrient information in spectral data can be revealed through calibration modeling with a regression approach. There are several available and commonly used techniques in NIR modeling that are based on linear and non-linear algorithms. The two most common calibration methods are principal component regression (PCR) and partial least square regression (PLSR) [3], [4]. PCR and PLSR models for prediction of soil N, P, K, Ph, Mg and Ca, PLSR performance is better than PCR [5]. The results of another literature review on the detection of N (Figure 1), P (Figure 2), and K (Figure 3) show that most of them use PLSR. The concept of PLSR is that the spectra data of soil samples (variable X) and soil elements (variable Y) are projected into a new space, orthogonal bases of latent variables are built one by one in such a way that they are oriented along the direction of maximum covariance between the spectra matrix and the response vector.

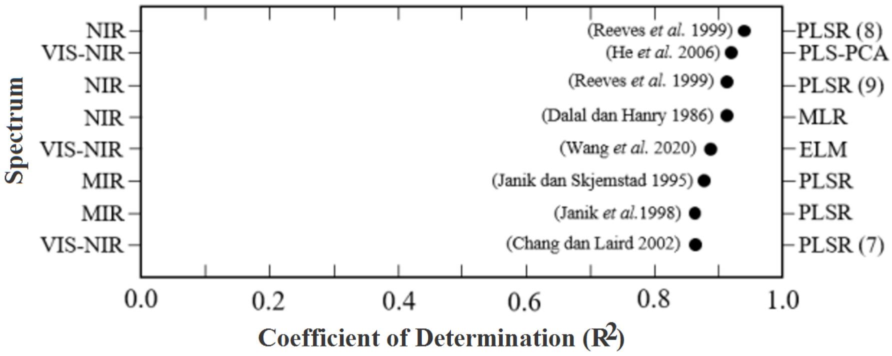

Figure 1. Diagram of spectroscopic literature review on soil nitrogen

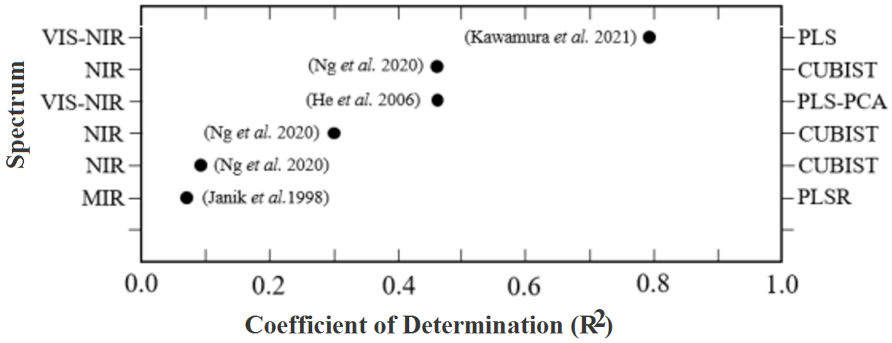

Figure 2. Diagram of spectroscopic literature review on soil Phosphorus element

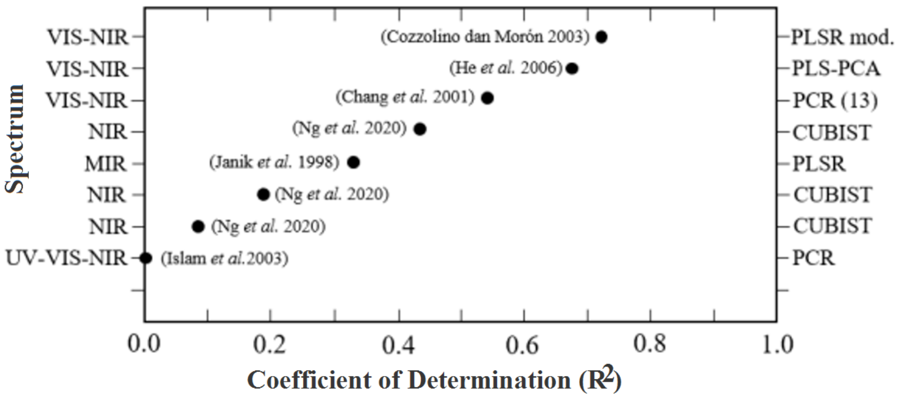

Figure 3. Diagram of spectroscopic literature review on soil Kalium

NIR measurements contribute to the presence of noise. Preprocessing is an important step in NIR data processing as it can improve model performance. The application of NIR spectroscopy is challenging because the recorded spectra contain a mixture of light absorption and light scattering effects [6]. Preprocessing aims to remove uninformative sources of variance (such as, scattering and instrumentation effects) from the measured spectra. The preprocessing step usually starts from data visualization and removal of extreme spectra dominated by noise. The next step performs a smoothing operation to remove high-frequency noise, a common method used is Savitzky-Golay (SAVGOL) which involves the use of polynomials of a chosen order of a certain size [7]. Ideally, in the presence of absorbance features, the smoothed spectra are ready for regression modeling or classification. However, due to the dominance of scattering effects, the smoothing step is usually followed by a scatter correction method. A commonly used operation is the standard normal variate (SNV) with the treatment of each spectra subtracting the average intensity of the spectra and then dividing it by the standard deviation estimated in the spectra domain [8]. Besides, there are also several other preprocessing methods and operations [9].

Research on the best preprocessing methods is often determined through trial-and-error. A more effective approach to optimizing preprocessing in NIR models is to compare a large number of preprocessing techniques. Several studies compare different preprocessing methods to produce optimal model inputs [10]. Choosing the right preprocessing technique is always a challenge. The effectiveness of different combinations of preprocessing methods has been investigated. The selection strategy and preprocessing experiments show different model performance [11]. A design-of-experiments approach to select the best preprocessing strategy can improve model performance and interpretation [12]. The use of genetic algorithms to find the optimal preprocessing strategy in Raman spectroscopy [13]. The sequence of preprocessing applied can have an effect on model performance [14], therefore it is necessary to find a good combination of preprocessing methods in a sequence, but two or more scatter correction operators are never used in a complementary way.

The development of a prediction model can be divided into three stages. The first stage, spectra preprocessing, aims to remove all uninformative sources of variance from the spectrum [14]. The second stage, feature selection, aims to select effective waves. Variable selection methods include various decomposition methods, sequential methods and optimization methods [15]. Information irrelevant to the sampled components in the original waveform can reduce the predictive ability of the model [16]. The third stage, calibration, trains a regression model that maps the extracted spectral features to describe the desired sample properties. Machine learning (ML) has the ability to make accurate predictions and refine models based on previous experience, being able to discover patterns and relationships that may not be visible to humans. By automatically processing data, ML can discover hidden patterns. ML models can learn from new data and automatically improve themselves to increase accuracy and efficiency. Some research in NIR data processing has better performance than PLS. The application of artificial neural network (ANN)/ Multi Layer Perceptron (MLP) and support vector machine (SVM) is better than PLSR for acidity prediction in mango fruit [17]. The application of random forest (RF) outperformed PLS performance for the prediction of soil alkaline and organic content [18]. Research on mango quality and maturity detection discusses the application of various ML algorithms, namely RF, SVM, MLP, LR, and PLS [19]. The use of these algorithms is intended to explore optimal classification performance based on spectrum data obtained from NIR spectroscopy devices.

This research proposes 2 new approaches to generate prediction performance, namely (1) Application of ML to find the best combination of preprocessing operations quickly and simultaneously so as to find an effective preprocessing strategy with hyperparameter tuning to find an optimized model. This approach will contribute to the preprocessing strategy because in the selection of operations there is no standardized standard and is still trial and error (2). The application of ML algorithms are RF, SVM, MLP, LR, and PLS. To find the best performance perform hyperparameter tuning on each algorithm. The results of the comparison of algorithms for prediction modeling will contribute to the soil nutrient prediction model that has been mostly using PLSR.

2. Research Method

This research generally starts from identifying problems and literature review, then continues the dataset collection stage as well as dataset analysis, then preprocessing is carried out which is an important step before model development. Model development by comparing 5 algorithms, namely 2 common algorithms and 3 new algorithms. The stages that will be worked on can be seen in Figure 4.

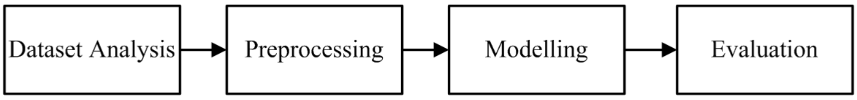

Figure 4. Research stages

Problem identification and literature review have been carried out by examining several studies related to soil fertility prediction using NIR spectroscopy and the use of machine learning. Until the development of the model, it was decided to use an open access dataset, namely NIR spectra datasets of agricultural soil and fertility properties [20]. NIR dataset with a wavelength range of 1000-2500 nm with a total of 40 samples. Soil fertility properties were measured by chemical analysis of soil nitrogen (N), phosphorus (P), potassium (K), soil pH, magnesium (Mg) and calcium (Ca).

Dataset analysis stage

This stage will look at the spectral peaks that affect each soil fertility element before building the model. The spectral peaks are interpreted as wavelengths that are sensitive to changes in that particular nutrient.

Preprocessing stage

Preprocessing techniques can improve model performance and interpretability [19]. There are several preprocessing methods, including clipping, scatter correction, smoothing, derivatives, trimming and resampling. The clipping method removes or replaces data points with specific values, eliminating noise that distorts information. Scatter correction methods aim to counter the effects of particle size. Some preprocessing operations related to the scatter correction method include SNV, MSC and normalization. SNV is able to remove additive and multiplicative effects due to light scattering, by involving the treatment of each spectral by subtracting the average spectral intensity from each intensity response and then dividing it by the standard deviation in the spectral domain. The non-parametric version of SNV is robust normal variate (RNV), RNV is more suitable for data with more noise, the correction concept is based on median values and inter-quartile intervals. The local version of SNV is called LSNV, as the SNV operation is performed piecewise in a spectral window. The MSC operation changes the spectral mean, the process is performed several times until the spectral mean no longer changes. The extended version of MSC is known as EMSC, which takes into account both linear and quadratic terms when performing the correction process. Spectral normalization can be performed over a range of 0 and 1 values, if no normalization range is provided, each spectral is normalized using Euclidean. The smoothing method aims to remove noise from the environment or instrumentation. Smooothing applies the Savitzky-Golay filter which is able to restore the smooth derivative of the original spectral. The trimming method allows the extraction of continuous and non-continuous spectral regions and resampling processes new spectral resolution using the fourier method which can combine spectral. This study uses 5 methods and 12 operations with details of operation details, parameters and values in Table 1. In improving the accuracy performance, preprocessing is carried out using 12 spectral transformation operators, then the operators will be collected and combined to obtain optimal performance [21]. Figure 5 shows the spectral transformation process and the best operation that will be used for the regression modeling stage.

Table 1. Methods, operations, parameters and values of spectral transforms

Methods	Operator	Parameter	Values
Clipping	CLIP		
Scatter Correction	SNV		
	RNV	iqr	75-25, 90-10
	LSNV		
	MSC		
	EMSC		
	NORML		
	BASELINE		

	SMOOTH	filter_win window_type	7, 11, 61 hamming
Derivative	DETREND	bp	0
Savinky Golay	SAVGOL	filter_win poly_order deriv_order	7, 11, 21, 61, 121 3, 6 0, 1, 2
Resampling	RESAMPLE	rasio	0.7

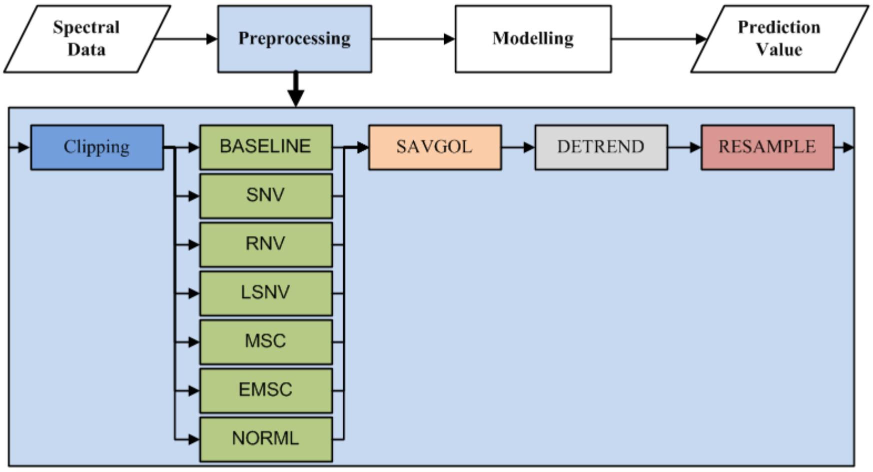

Figure 5. The operation principle of collecting and combining spectral transform operators

Machine learning will perform automated learning and hyperparameter tuning methods to find further optimized models. The effect of different parameter values in the operator on model performance has rarely been studied in depth. Using machine learning will test various preprocessing combinations quickly and simultaneously, so as to find effective preprocessing strategies.

Modelling stage

Data processing is done using scikit-learn, keras, and tensorflow libraries with python programming language. The dataset was divided into training and testing data with a proportion of 90:10. The dataset was then trained with machine learning algorithms namely RF, SVM, MLP, LR, and PLS. The performance of the models is compared with each other and the best one is concluded.

Evaluation stage

Model evaluation is done by looking at the Mean Square Error (MSE) and root mean square error (RMSE) values. The best prediction results when the model has the least error rate. MSE is the average squared error between the actual value and the forecasting value. MSE is generally used to check the estimation of how much error is in the prediction. A low MSE value or close to zero indicates that the prediction results are in accordance with the actual data and can be used for forecasting calculations in the future period. The way to calculate MSE is to subtract the actual data value from the predicted data and the results are squared, then summed up as a whole and divided by the amount of data available. RMSE is the magnitude of the prediction error rate, where the smaller or closer to 0 the RMSE value, the more accurate the prediction results will be. RMSE is one way to evaluate linear regression models by measuring the accuracy of a model's forecast results. RMSE has no units. A low RMSE value indicates that the variation in values produced by a forecast model is close to the variation in observed values. RMSE calculates how different a set of values are. The smaller the RMSE value, the closer the predicted and observed values are.

3. Literature Study

3.1. Near Infrared Spectroscopy

Near infrared (NIR) spectroscopy is an analytical technique to determine the chemical composition or structure of a particular sample. Spectroscopy is a science that discusses the interaction of light with molecules and atoms [22]. The level of absorbance of energy by the sample will be captured by the detector on the spectrometer according to the electromagnetic wave region. The NIR covers a wide range of the electromagnetic spectrum between 780 nm and 2500 nm. The captured NIR spectra of biological objects consist of the response of O-H, C-H, C-O and N-H molecular bonds. These bonds are subject to vibrational energy changes when illuminated by NIR frequencies [23]. The target sample is illuminated with NIR and the reflected and backscattered light is measured with a spectrometer. The NIR-active molecular bonds in the sample absorb incoming light at different tonal spectral bands and spectrum combinations, resulting in an NIR absorbance spectrum.

3.2. Preprocessing Spectral

NIR spectral has hundreds or thousands of wavelengths, the developed prediction model will become too complicated if all spectral wavelengths are used directly. In addition, some irrelevant information including noise and stray light caused by the instrument or environment affect the identification of spectrum information in the modeling process, which may weaken the predictive ability of the developed model [24]. Preprocessing methods are helpful for developing reliable models by removing the interference of irrelevant information [25]. Preprocessing techniques can improve model performance and interpretability [26]. The techniques include 6 categories including: Clipping, Scatter Correction, Smoothing, Derivatives, Trimming and Resampling. The order of preprocessing operations applied can have an effect on model performance [14]. To make the calibration model determine how well the technique performs, a large amount of research has focused on optimizing this process. Attempts to compare different preprocessing operations to obtain the best model input [10]. A critical review of preprocessing strategies showed that model performance between different preprocessing strategies [7]. It should be noted, that the order in which preprocessing operations are performed can have an effect on model performance.

3.3. Machine Learning Modelling for NIR Spectral

Machine Learning (ML) serves to improve data processing and analysis in NIR spectroscopy by utilizing algorithms to handle large and complex data. There are 3 challenges in modeling NIR data namely (1) High Dimensionality, NIR spectroscopy data has thousands of data points as it covers many wavelengths. This can cause challenges in processing and analysis. (2) Noise and Variability, NIR data is often affected by noise and variability from the sample which requires specialized processing techniques. (3) Non-Linear Relationships, the relationship between NIR spectra and analysis properties is often non-linear, requiring analysis methods that can handle this complexity [27]. Some of the algorithms used include, Random Forest (RF) is an ensemble algorithm that combines several decision trees to produce more accurate and stable predictions. RF in its work includes 3 parts, namely bootstrapping, feature randomization, and averaging/voting. RF has the advantage of being robust to overfitting and handling high-dimensional data. Support Vector Regression (SVR) is a regression method for predicting continuous values. SVR attempts to find a function that minimizes the prediction error while still having a small margin of error. SVR works based on error margins and kernel tricks. the advantages of SVR are the ability to handle non-linear data and stable performance. Partial Least Squares (PLS) is a regression technique that combines elements of principal component analysis (PCA) and regression. PLS works with the principles of dimensionality reduction and regression. PLS is able to overcome multicollinearity and increase interpretability. Linear Regression (LR) is a statistical method for modeling the relationship between a dependent variable and one or more independent variables. This model attempts to find a hyperplane by working with a linear model and parameter estimation. The advantages of LR are simplicity and interpretability as well as fast and efficient computation [28]. Multi-Layer Perceptron (MLP) is a type of neural network consisting of several layers of neurons. The way MLP works is based on layers of neurons and backpropagation. The advantages of MLP are that it can capture non-linear relationships and is flexible [29].

4. Result and Discussion

4.1. Raw Spectrum Analysis

The spectrum acquisition resulted in a spectrum of the raw data, where the spectrum can be seen in Figure 6. Based on Figure 6, there are relevant wavelengths for chemical bonds that describe soil properties and elements. Each material has different optical characteristics and electromagnetic spectrum shapes, where the shape of this spectrum will characterize the chemical content of the material. The raw spectrum shows noise caused by interference during data collection. Therefore, it is necessary to improve the raw data spectrum to reduce noise.

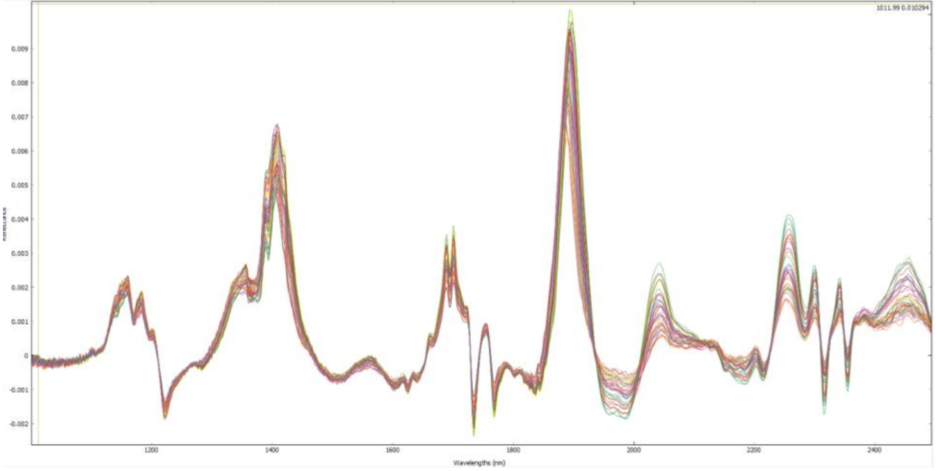

Figure 6. Raw spectrum data of soil samples

4.2. Modeling Without Preprocessing

Modeling results without preprocessing, namely from direct spectrum data modeled using the five algorithms namely RF, SVR, PLS, LR, and MLP which are evaluated using MSE and RMSE values can be seen in Table 2.

Table 2. Modeling results without preprocessing

NONE	RF		SVR		PLS		LR		MLP	
	MSE	RMSE	MSE	RMSE	MSE	RMSE	MSE	RMSE	MSE	RMSE
K	0.275	0.524	0.299	0.547	0.256	0.506	0.240	0.490	0.275	0.524
Mg	23.408	4.838	23.447	4.842	34.352	5.861	39.957	6.321	23.282	4.825
Ca	46.752	6.838	48.166	6.940	73.892	8.596	71.218	8.439	43.647	6.607
P	133.686	11.562	158.937	12.607	160.504	12.669	128.498	11.336	135.421	11.637
pH	5.623	2.371	5.361	2.315	6.662	2.581	5.014	2.239	4.870	2.207
N	0.017	0.132	0.021	0.143	0.022	0.148	0.024	0.154	0.023	0.150

4.3. Modeling Using Preprocessing

Modeling results using preprocessing with the best combination of 12 operators, parameters and their values are shown in Table 3. The MSE and RMSE values of the five algorithms are shown and it can be seen that there is a decrease in error or an increase in performance of each model used.

Table 3. Modeling results using preprocessing

Prepro-cessing	RF		SVR		PLS		LR		MLP	
	MSE	RMSE	MSE	RMSE	MSE	RMSE	MSE	RMSE	MSE	RMSE
K	0,176	0,420	0,250	0,500	0,205	0,452	0,156	0,394	0,244	0,494
Mg	3,117	1,766	8,386	2,896	5,673	2,382	7,469	2,733	2,761	1,662
Ca	14,574	3,818	37,190	6,098	23,464	4,844	20,447	4,522	11,889	3,448
P	10,900	3,301	28,465	5,335	45,494	6,745	25,852	5,084	9,996	3,162
pH	3,137	1,771	2,447	1,564	4,532	2,129	1,173	1,083	0,351	0,593
N	0,004	0,063	0,009	0,094	0,014	0,120	0,009	0,095	0,006	0,081

4.4. Modelling Comparison

The use of preprocessing techniques can reduce the error, this can be seen in all algorithms in testing all soil elements (K, Mg, Ca, P, pH, and N) in Figure 7 and Figure 8. The error value with the use of preprocessing techniques is much smaller than without using preprocessing. A significant decrease in error values in all algorithms is seen in testing soil elements P and Mg.

Comparison of the five algorithms in testing the six soil elements (K, Mg, Ca, P, pH, and N) that all algorithms succeeded in reducing the error, can be seen in Figure 9 and Figure 10. Comparison of the five algorithms on the K soil element that has the lowest error is using LR, while in testing the elements of Mg, Ca, P, and pH the MLP algorithm has good performance, in testing the N soil element the best performance is the RF algorithm followed by the MLP algorithm.

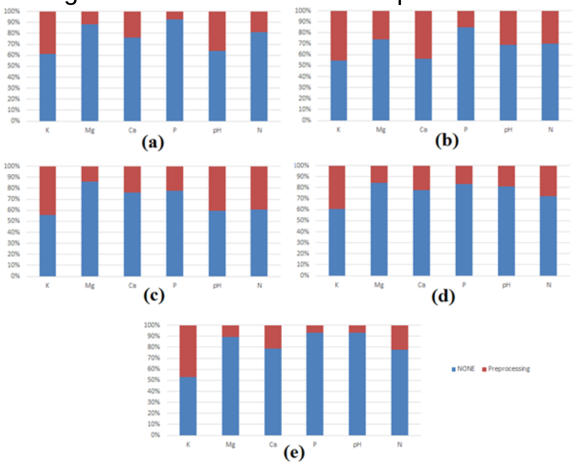

Figure 7. Comparison of MSE on (a) RF, (b) SVR, (c) PLS, (d) LR, (e) MLP

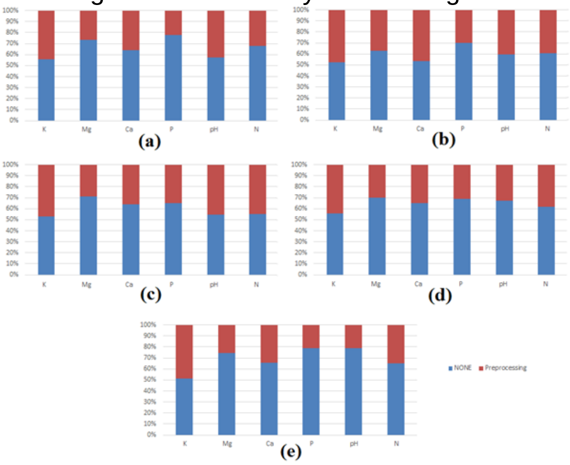

Figure 8. Comparison of RMSE on (a) RF, (b) SVR, (c) PLS, (d) LR, (e) MLP

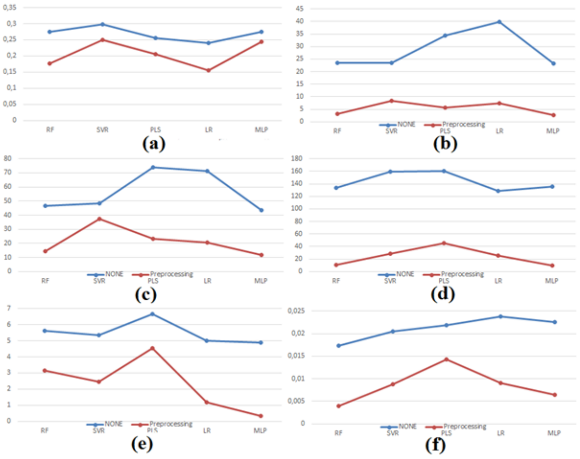

Figure 9. Comparison of MSE on (a) K, (b) Mg, (c) Ca, (d) P, (e) pH, (f) N

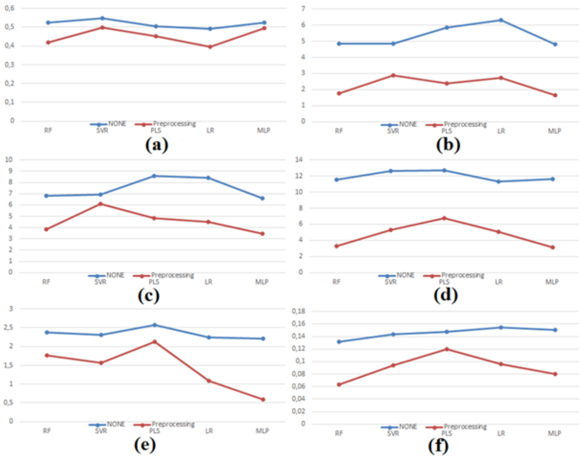

Figure 10. Comparison of RMSE on (a) K, (b) Mg, (c) Ca, (d) P, (e) pH, (f) N

The algorithm with the best performance on each soil element predictor (K, Mg, Ca, P, pH, and N) uses preprocessing techniques. The combination of operators, parameters, and values of preprocessing in each algorithm can be seen in Table 4. The LR algorithm for soil element K uses a combination of CLIP, LSNV, RESAMPLE, and SMOOTH operators. The combination of operators in preprocessing for predicting the soil element Mg using the MLP algorithm is CLIP, DETREND, LSNV, RESAMPLE, and SMOOTH. The combination of operators in preprocessing for soil elements Ca, P, pH and N in more detail can be seen in Table 4.

Table 4. Best preprocessing combination according to soil elements and algorithm

Soil	Algoritm	Operator	Parameter and value
K	LR	CLIP	{'substitute': None, 'threshold': 10000.0}

		LSNV	{}
		RESAMPLE	{'resampling_ratio': 0.7}
		SMOOTH	{'filter_win': 61, 'window_type': 'hamming'}
Mg	MLP	CLIP	{'substitute': None, 'threshold': 10000.0}
		DETREND	{'bp': [0]}
		LSNV	{}
		RESAMPLE	{'resampling_ratio': 0.7}
		SMOOTH	{'filter_win': 11, 'window_type': 'hamming'}
Ca	MLP	CLIP	{'substitute': None, 'threshold': 10000.0}
		DETREND	{'bp': [0]}
		RNV	{'iqr': [75.0, 25.0]}
		SAVGOL	{'deriv_order': 0, 'filter_win': 121, 'poly_order': 3}
P	MLP	CLIP	{'substitute': None, 'threshold': 10000.0}
		DETREND	{'bp': [0]}
		RNV	{'iqr': [90.0, 10.0]}
		SMOOTH	{'filter_win': 61, 'window_type': 'hamming'}
pH	MLP	CLIP	{'substitute': None, 'threshold': 10000.0}
		DETREND	{'bp': [0]}
		LSNV	{}
		SAVGOL	{'deriv_order': 1, 'filter_win': 61, 'poly_order': 6}
N	RF	CLIP	{'substitute': None, 'threshold': 10000.0}
		DETREND	{'bp': [0]}
		EMSC	{}
		RESAMPLE	{'resampling_ratio': 0.7}
		SAVGOL	{'deriv_order': 2, 'filter_win': 61, 'poly_order': 3}

5. Conclusion

Spectrum preprocessing is an important step in NIR processing as it can improve model performance. The use of the best spectral transformation method is often determined through trial and error. Testing the most optimal preprocessing operation, all operators were collected and combined to obtain optimal performance by measuring the performance of each algorithm by measuring the MSE and RMSE values. Test results with preprocessing using 12 operators resulted in 2,112 operator combinations. The use of preprocessing techniques can improve the performance of all algorithms (RF, SVR, PLS, LR, and MLP). Testing the soil element K which has the lowest error is using LR, in testing the soil elements Mg, Ca, P, and pH using the MLP algorithm has the best performance and in testing the soil element N the best performance on the RF algorithm.

Acknowledgement

The authors would like to thank the Directorate of Research, Technology and Community Service (DRTPM) of the Ministry of Education and Culture of the Republic of Indonesia for supporting and funding research in 2023.

References

[1] Sanseechan P, Panduangnate L, Saengprachatanarug K, Wongpichet S, Taira E, and Posom J. A portable near infrared spectrometer as a non-destructive tool for rapid screening of solid density stalk in a sugarcane breeding program. Sens. Bio-Sensing Res., vol. 20, pp. 34–40, Sep. 2018, doi: 10.1016/j.sbsr.2018.07.001.

[2] Mohamed ES, Saleh AM, Belal AB, and Gad A. Application of near-infrared reflectance for quantitative assessment of soil properties. Egypt. J. Remote Sens. Sp. Sci., vol. 21, no. 1, pp. 1–14, Apr. 2018, doi: 10.1016/j.ejrs.2017.02.001.

[3] Munawar AA, Kusumiyati, and Wahyuni D. 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.

[4] Agussabti, Rahmaddiansyah, Satriyo P, and Munawar AA. Data analysis on near infrared spectroscopy as a part of technology adoption for cocoa farmer in Aceh Province, Indonesia. Data Br., vol. 29, p. 105251, Apr. 2020, doi: 10.1016/j.dib.2020.105251.

[5] Yunus Y, Devianti, Satriyo P, and Munawar AA. Rapid Prediction of Soil Quality Indices Using Near Infrared Spectroscopy. IOP Conf. Ser. Earth Environ. Sci., vol. 365, no. 1, p. 012043, Oct. 2019, doi: 10.1088/1755-1315/365/1/012043.

[6] Lu R, Van Beers R, Saeys W, Li C, and Cen H. Measurement of optical properties of fruits and vegetables: A review. Postharvest Biol. Technol., vol. 159, p. 111003, Jan. 2020, doi: 10.1016/j.postharvbio.2019.111003.

[7] Rinnan A, van den Berg F, and Engelsen SB. Review of the most common pre-processing techniques for near-infrared spectra. TrAC Trends Anal. Chem., vol. 28, no. 10, pp. 1201–1222, Nov. 2009, doi: 10.1016/j.trac.2009.07.007.

[8] Barnes RJ, Dhanoa MS, and Lister SJ. Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra. Appl. Spectrosc., vol. 43, no. 5, pp. 772–777, Jul. 1989, doi: 10.1366/0003702894202201.

[9] Roger JM, Boulet JC, Zeaiter M, and Rutledge DN. Pre-processing Methods. in Comprehensive Chemometrics, Elsevier, 2020, pp. 1–75.

[10] Xu L. Ensemble preprocessing of near-infrared (NIR) spectra for multivariate calibration. Anal. Chim. Acta, vol. 616, no. 2, pp. 138–143, Jun. 2008, doi: 10.1016/j.aca.2008.04.031.

[11] Engel J. Breaking with trends in pre-processing?. TrAC Trends Anal. Chem., vol. 50, pp. 96–106, Oct. 2013, doi: 10.1016/j.trac.2013.04.015.

[12] Gerretzen J. Boosting model performance and interpretation by entangling preprocessing selection and variable selection. Anal. Chim. Acta, vol. 938, pp. 44–52, Sep. 2016, doi: 10.1016/j.aca.2016.08.022.

[13] Bocklitz T, Walter A, Hartmann K, Rösch P, and Popp J. How to pre-process Raman spectra for reliable and stable models?. Anal. Chim. Acta, vol. 704, no. 1–2, pp. 47–56, Oct. 2011, doi: 10.1016/j.aca.2011.06.043.

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

[15] Balabin RM and Smirnov SV. Variable selection in near-infrared spectroscopy: Benchmarking of feature selection methods on biodiesel data. Anal. Chim. Acta, vol. 692, no. 1–2, pp. 63–72, Apr. 2011, doi: 10.1016/j.aca.2011.03.006.

[16] Xu H, Qi B, Sun T, Fu X, and Ying Y. Variable selection in visible and near-infrared spectra: Application to on-line determination of sugar content in pears. J. Food Eng., vol. 109, no. 1, pp. 142–147, Mar. 2012, doi: 10.1016/j.jfoodeng.2011.09.022.

[17] Munawar AA, Zulfahrizal, Meilina H, and Pawelzik E. Near infrared spectroscopy as a fast and non-destructive technique for total acidity prediction of intact mango: Comparison among regression approaches. Comput. Electron. Agric., vol. 193, p. 106657, Feb. 2022, doi: 10.1016/j.compag.2021.106657.

[18] Santana FB,de Souza AM, and Poppi RJ. Visible and near infrared spectroscopy coupled to random forest to quantify some soil quality parameters. Spectrochim. Acta Part A Mol. Biomol. Spectrosc., vol. 191, pp. 454–462, Feb. 2018, doi: 10.1016/j.saa.2017.10.052.

[19] Khumaidi A, Purwanto YA, Sukoco H, and Wijaya SH. Using Fuzzy Logic to Increase Accuracy in Mango Maturity Index Classification: Approach for Developing a Portable Near-Infrared Spectroscopy Device. Sensors, vol. 22, no. 24, p. 9704, Dec. 2022, doi: 10.3390/s22249704.

[20] Munawar AA, Yunus Y, and Devianti D. NIR spectra datasets of agricultural soil and fertility properties. Mendeley Data, vol. IV, 2020, doi: 10.17632/h8mht3jsbz.1.

[21] Khumaidi A, Purwanto YA, Sukoco H, and Wijaya SH. Effects of spectral transformations in support vector machine on predicting ‘Arumanis’ mango ripeness using near-infrared spectroscopy. J. Ilm. Ilk., vol. 13, no. 3, 2021.

[22] C. J. Clifford, R. A. Olaf, and C. M. Malcolm, *Analisis Spektrum Senyawa Organik*. Bandung: Institut Teknologi Bandung, 2005.

[23] K. B. Walsh, V. A. McGlone, and D. H. Han, “The uses of near infra-red spectroscopy in postharvest decision support: A review,” *Postharvest Biol. Technol.*, vol. 163, p. 111139, May 2020, doi: 10.1016/j.postharvbio.2020.111139.

[24] D. Zhang *et al.*, “Nondestructive measurement of soluble solids content in apple using near infrared hyperspectral imaging coupled with wavelength selection algorithm,” *Infrared Phys. Technol.*, vol. 98, pp. 297–304, May 2019, doi: 10.1016/j.infrared.2019.03.026.

[25] L. S. Magwaza and U. L. Opara, “Analytical methods for determination of sugars and sweetness of horticultural products—A review,” *Sci. Hortic. (Amsterdam).*, vol. 184, pp. 179–192, 2015, doi: https://doi.org/10.1016/j.scienta.2015.01.001.

[26] J. Gerretzen *et al.*, “Boosting model performance and interpretation by entangling preprocessing selection and variable selection,” *Anal. Chim. Acta*, vol. 938, pp. 44–52, Sep. 2016, doi: 10.1016/j.aca.2016.08.022.

[27] Y. Xu, J. Ge, and C.-W. Ju, “Machine learning in energy chemistry: introduction, challenges and perspectives,” *Energy Adv.*, vol. 2, no. 7, pp. 896–921, 2023, doi: 10.1039/D3YA00057E.

[28] J. R. Rokde and A. G. Thosar, “Linear regression approach for performance evaluation of ES with load impedance variations of non-critical and critical load,” *e-Prime - Adv. Electr. Eng. Electron. Energy*, vol. 6, p. 100312, Dec. 2023, doi: 10.1016/j.prime.2023.100312.

[29] J. Liu, X. Bai, L. Zhang, R. Cai, and M. Hu, “Influencing factors for water use behaviors of population groups in hospitals based on multi-layer perceptron model,” *Desalin. Water Treat.*, vol. 290, pp. 185–192, Apr. 2023, doi: 10.5004/dwt.2023.29422.