



Classification of Impurity Levels in Contaminated Sugarcane Leaf Pellets using NIR Spectroscopy

Thanaporn Saksuwan¹, Waraporn Sawangchom¹, Jetsada Posom^{1,3} and Kanvisit Maraphum^{2*}

¹Department of Agricultural Engineering, Faculty of Engineering, Khon Kaen University, Khon Kaen, 40002, Thailand.

²Department of Agricultural Machinery, Faculty of Agriculture and Technology, Rajamangala University of Technology Isan Surin Campus, Surin, 32000, Thailand

³Center for Alternative Energy Research and Development, Khon Kaen University, Khon Kaen, 40002, Thailand

*Corresponding author.

Email address: kanvisit.ma@rmuti.ac.th

doi: 10.14456/abe.2026.9

Received 31 March 2026

Revised 10 June 2026

Accepted 23 June 2026

Abstract

This research investigates the feasibility of classifying impurity levels in sugarcane leaf pellets. Samples were prepared by mixing sugarcane leaves with varying proportions of sand. NIR spectroscopy combined with linear discriminant analysis (LDA) was employed to categorize samples across discrete contamination tiers (5% to 25% sand content). Because an absolute 0% clean baseline was not evaluated, the framework is interpreted strictly as a system for grading impurity severity levels. Experimental results revealed that centering and 1st derivative preprocessing yielded the highest accuracy. These techniques effectively mitigated interference from light scattering and baseline shifts, while enhancing specific spectral features related to biomass chemical constituents, such as cellulose, lignin, and hemicellulose. The LDA model, optimized with preprocessing, achieved classification accuracy exceeding 97%, with minimal errors observed in samples containing low to moderate sand content. However, accuracy slightly declined when sand content exceeded 20% due to reflective interference from silica (SiO₂). The findings confirm that selecting appropriate preprocessing methods is a critical factor for model performance. These laboratory-scale results demonstrate the strong potential of the method, serving as a foundational step toward developing future on-line monitoring pipelines for real-time soil contamination tracking.

Keywords: Sugarcane leaf pellets, Renewable energy, Near-infrared spectroscopy (NIRS), trash

1. Introduction

Due to the upward trend in national energy consumption, Thailand has set a target to increase the proportion of renewable and alternative energy to 30% of total energy consumption by 2037 [1]. Post-harvest sugarcane leaf residue has the potential to serve as biomass fuel, capable of generating over 400 MW of electricity. Currently, sugarcane leaves are traded as fuel for biomass power plants in the form of leaf rolls or bales, priced between 500 and 1,000 baht/T [2]. When considering the total value, this could generate up to 23 billion baht in income for sugarcane farmers and the sugar industry from over 10 million tons of leaf residue [3]. However, the sale of baled sugarcane leaves is currently limited to large-scale farmers, accounting for only 40% of the total area where sugarcane burning is avoided. In contrast, 60% of small-scale farmers still engage in sugarcane burning. If this portion of leaf residue could be utilized, it would significantly increase value-added opportunities and expand the production capacity of biomass fuel from agricultural waste.

Collecting sugarcane leaves in the form of rolls or large bales poses significant logistical and management challenges. A 10-wheel truck can carry only 10 tons, accommodating merely 15–20 bales depending on their size. Furthermore, transporting sugarcane leaves beyond a 30-km radius from the purchasing plant is not cost-effective for farmers. Increasing the density of sugarcane leaves by processing them into pellets can optimize storage space. Additionally, the purchase price for sugarcane leaf pellets at power plants can rise to 3,000–3,500 baht/T [4], while also reducing the risk of spontaneous combustion.

However, the production of pellets faces challenges regarding impurities, particularly the contamination of soil and sand within the sugarcane leaf compacts. The presence of such contaminants in pellets adversely affects the combustion system and the thermal energy output. Currently, near infrared (NIR) spectroscopy technology is employed for the precise, real-time inspection and quality assessment of biomass pellets in agricultural production. This method can be performed using both contact and non-contact techniques, such as predicting moisture content and temperature of biomass pellets [5], evaluating the pyrolysis degradation characteristics of ground and pelletized bamboo [6], and assessing proximate data including moisture, ash, volatile matter, fixed carbon, and lignocellulose content

(cellulose, hemicellulose, and lignin) of corn cobs [7]. Moreover, NIR has been used to predict moisture, heating value, ash content, and carbon in energy crops like Miscanthus and Short Rotational Coppice Willow (SRCW) [8], as well as determining the moisture, carbon, ash content, and gross calorific value (GCV) of biomass pellets [9]. This technology has been proven to be a rapid and accurate predictive method, enabling effective product quality control and improvement [10].

Therefore, this research aims to evaluate the amount of sand contamination in sugarcane leaf pellets using NIR data combined with machine learning algorithms.

2. Materials and methods

2.1 Sample Preparation

The sugarcane leaves were shredded into small pieces with a maximum length of 5 cm. These shredded leaves were then used to produce pellets, categorized into 16 sample groups. The sand used as an impurity agent was standard river sand that was thoroughly washed, oven-dried at 105°C for 24 h to eliminate moisture variables and sieved to isolate a controlled particle size range of 0.25–0.50 mm. Maintaining a consistent sand grade and particle size is critical, as grain morphology directly influences diffuse reflectance boundaries and silica-induced light-scattering behavior under NIR inspection. The percentage of impurities (sand) was varied among the groups as detailed in Table 1, specifically at levels of 5%, 10%, 15%, 20%, and 25% by weight. The sand and shredded leaves were thoroughly mixed before being compressed into cylindrical pellets, each with a diameter of 5 cm and a height of 5 cm. A total of 16 samples were produced, representing different impurity levels, as illustrated in Figure 1.

Table 1 Percentage of impurities in sugarcane leaf pellet samples

Sample	Sand weight (g)	Sugarcane leaf weight (g)	Total weight (g)	Contaminate (%)
1	2.5	50	52.5	5
2	2.5	50	52.5	5
3	2.5	50	52.5	5
4	2.5	50	52.5	5
5	5.0	50	55	10
6	5.0	50	55	10
7	5.0	50	55	10
8	7.5	50	57.5	15
9	7.5	50	57.5	15
10	7.5	50	57.5	15
11	10.0	50	60	20
12	10.0	50	60	20
13	10.0	50	60	20
14	12.5	50	62.5	25
15	12.5	50	62.5	25
16	12.5	50	62.5	25



Figure 1 Sugarcane leaf pellet samples with varying levels of impurities (sand)

2.2 Spectral Acquisition

The samples were scanned using a NIR spectrometer across the wavelength range of 860–1760 nm (AvaSpec-NIR256/512–1.7-EVO, Avantes BV, Apeldoorn, Netherlands). Spectral data were collected using a fiber optic probe and an integrating sphere. To account for surface variations and non-homogeneous sand distribution within the matrix, multiple optical measurements were taken per sample. Specifically, for each individual physical pellet, 30 separate NIR spectral scans were acquired. Between each sequential scan, the fiber-optic probe was lifted, the sample was rotated, and the probe was repositioned at a different point on the sample surface, producing a total of 480 raw spectra for the entire dataset. Between each sequential scan, the fiber-optic probe was lifted, the sample was rotated, and the probe was repositioned at a different point on the surface. This scanning procedure, illustrated in Figure 2. This was implemented to capture localized sample heterogeneity and provide a robust, averaged spectral profile for each physical sample. Crucially, these repeated scans were treated as correlated measurements rather than independent observations during model training and validation; grouping the spectral data by physical sample ID ensured that no individual sample scans spanned both the calibration and prediction sets simultaneously, thereby preventing optimistic model bias.



Figure 2 Spectral measurement setup using a fiber optic probe coupled with an integrating sphere

2.3 Classification Model Development

The classification models were developed using the LDA algorithm. LDA is a supervised machine learning and statistical technique designed to find a linear combination of independent variables that best separates two or more classes of objects.

The dataset was partitioned into a training set and a test set using a 70:30 ratio. To optimize model performance, various spectral pre-processing techniques were evaluated as experimental factors including Raw spectra, mean centering, Min-Max normalization, multiplicative scatter correction (MSC), 1st derivative (Savitzky-Golay), 2nd derivative (Savitzky-Golay), standard normal variate (SNV), Scaling (Auto-scaling) and baseline offset correction, respectively.

Then, each pre-processing technique was applied to the spectral data before training the LDA models. For the first and second derivative transformations were executed using the Savitzky-Golay algorithm. To optimize the signal-to-noise ratio and prevent the inflation of high-frequency noise, the derivative parameters were strictly controlled with a moving window size of 15 points and a polynomial order of 2. The performance of each model was then compared based on classification accuracy to identify the most effective pre-treatment for detecting impurities in sugarcane leaf pellets.

2.4 Model Performance Evaluation

Note that the 5% sand contamination class contains 40 spectra while the other classes contain 30 spectra each. While this introduces a minor class imbalance, its mathematical weight on the macro-averaged performance metrics was found to be negligible; nonetheless, it is considered during per-class metric interpretation.

The dataset was partitioned using a strict sample-level split, assigning 11 physical pellets to the calibration set and 5 physical pellets to the independent test set. While 5 test pellets originally generated 150 spectral (5×30 scans), 4 spectral scans were identified as outliers during preprocessing and were removed, resulting in a final verified test set of 146 spectra, respectively.

A confusion matrix was employed as a tool to evaluate the accuracy of the classification models. It displays the number of correctly and incorrectly classified samples for each category, providing a detailed overview of the model's overall performance, as presented in Table 2. To evaluate model performance for the multi-class classification problem, classification performance metrics including true positive (TP), false positive (FP), false negative (FN), and true negative (TN) were calculated using a one-vs-rest macro-averaging approach. For any specific target class *I*, TP represents the number of spectra correctly classified into class *I*; FP represents the number of out-of-class spectra incorrectly assigned to class *I*; FN denotes the number of class *I* spectra incorrectly assigned to other categories; and TN represents the correct rejection of all non-class-*I* spectra, respectively. Once the TP, FP, FN, and TN values were obtained from each model, they are utilized to derive performance indicators shown in Table 3.

Table 2 Confusion matrix of the classification model

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

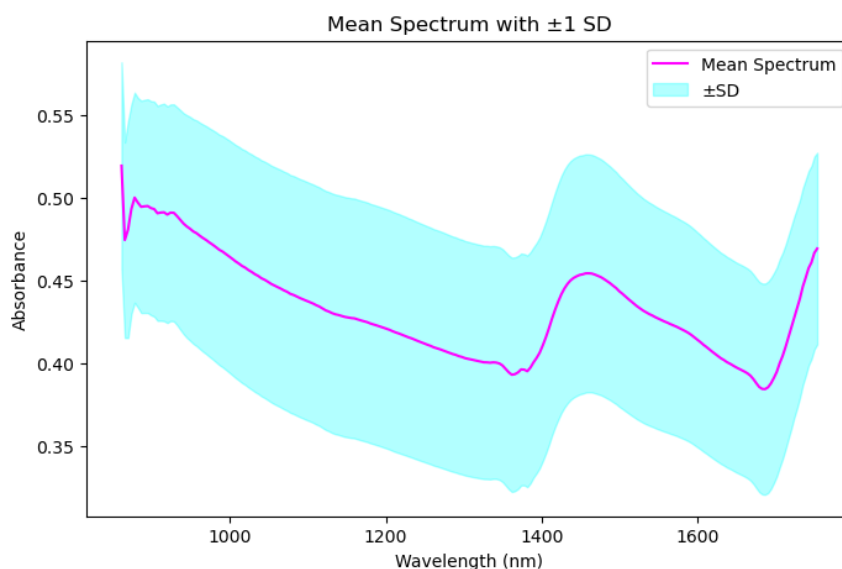
Table 3 Performance metrics of the classification models

Metric	Formula	Description
Accuracy	$(TP + TN) / (TP + TN + FP + FN)$	The percentage of total samples correctly classified by the model.
Precision	$TP / (TP + FP)$	The proportion of samples predicted as a specific class that were actually correct.
Recall / Sensitivity	$TP / (TP + FN)$	The proportion of actual positive samples that the model correctly identified.
Specificity	$TN / (TN + FP)$	The model's ability to correctly identify and exclude samples from other classes.
F1-score	$2 \times (Precision \times Recall) / (Precision + Recall)$	The harmonic mean of Precision and Recall, providing a balanced measure of performance.

3. Results and discussion

3.1 Spectra of sugarcane leaf pellets

Figure 3 illustrates the raw spectra of sugarcane leaf pellets with varying sand content ratios. It was observed that the spectra exhibit high variability across the wavelength range. A prominent absorption peak is evident at 1450 nm, which corresponds to the H₂O molecular vibrations. Figure 4 displays the spectra processed using different pre-treatment methods including (a) raw spectra, (b) mean centering, (c) min-max normalization, (d) MSC, (e) 1st derivative, (f) SNV, (g) scaling, and (h) baseline offset, respectively. Each spectral pre-processing technique aims to reduce the physical effects caused by variations in particle size. The results of these transformations, as shown in Figure 4, indicate that the pre-treatments significantly influence the spectral baseline and peak resolution. Notably, the derivative method effectively enhances hidden absorption peaks and mitigates the light-scattering interference.

**Figure 3** Raw spectra and standard deviation (SD) across the wavelength range of sugarcane leaf pellets.

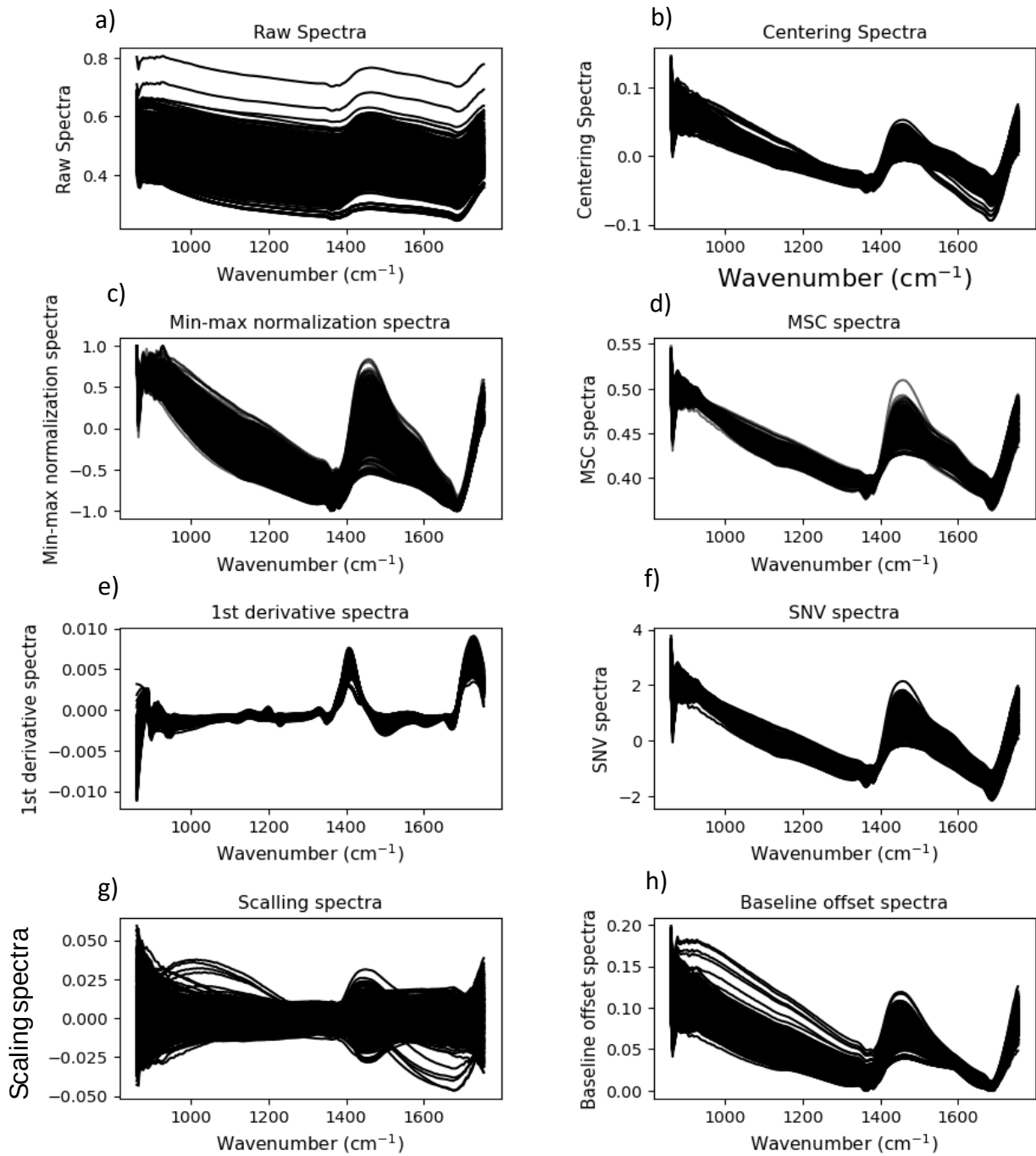


Figure 4 Spectral data of sugarcane leaf pellets processed with various pre-treatment methods including (a) raw spectra, (b) mean centering, (c) min-max normalization, (d) MSC, (e) 1st derivative, (f) SNV), (g) scaling, and (h) baseline offset

3.2 Modeling Results

The classification performance of the linear discriminant analysis (LDA) models varied depending on the pre-processing methods employed. Table 4 displayed the mean centering and scaling treatments that gave the best results, achieving a maximum Accuracy of 0.99 and an average F1-score exceeding 0.96 across various sand-mixture ratios.

The first and second derivatives provided the next best performance, with accuracy ranging from 0.95 to 0.97 and average F1-scores between 0.91 and 0.95. Meanwhile, MSC and SNV successfully mitigated light-scattering effects, their overall efficiency was slightly lower, with average Accuracies between 0.90 and 0.96. Meanwhile, Min-Max normalization and baseline offset showed moderate classification performance, with accuracy approximately between 0.89 and 0.94.

Table 4 Classification results of trash content percentages using the LDA algorithm

Pretreatment	Groups	N	TP	FN	FP	TN	Recall	Precision	Accuracy	F1-score
Raw spectra	5	146	39	2	1	104	0.95	0.98	0.98	0.96
	10	146	21	1	2	122	0.95	0.91	0.98	0.93
	15	146	21	3	5	117	0.88	0.81	0.95	0.84
	20	146	23	6	6	111	0.79	0.79	0.92	0.79
	25	146	27	3	1	115	0.90	0.96	0.97	0.93
Centering	5	146	39	2	0	105	0.95	1.00	0.99	0.98
	10	146	22	0	2	122	1.00	0.92	0.99	0.96
	15	146	21	3	5	117	0.88	0.81	0.95	0.84
	20	146	23	6	6	111	0.79	0.79	0.92	0.79
	25	146	27	3	1	115	0.90	0.96	0.97	0.93
Min_max normalization	5	146	37	4	2	103	0.90	0.95	0.96	0.93
	10	146	20	2	4	120	0.91	0.83	0.96	0.87
	15	146	20	4	7	115	0.83	0.74	0.92	0.78
	20	146	20	9	7	110	0.69	0.74	0.89	0.71
	25	146	27	3	2	114	0.90	0.93	0.97	0.92
MSC	5	146	37	4	1	104	0.90	0.97	0.97	0.94
	10	146	21	1	4	120	0.95	0.84	0.97	0.89
	15	146	20	4	7	115	0.83	0.74	0.92	0.78
	20	146	22	7	8	109	0.76	0.73	0.90	0.75
	25	146	25	5	1	115	0.83	0.96	0.96	0.89
1st derivative	5	146	38	3	1	104	0.93	0.97	0.97	0.95
	10	146	21	1	3	121	0.95	0.88	0.97	0.91
	15	146	20	4	3	119	0.83	0.87	0.95	0.85
	20	146	25	4	7	110	0.86	0.78	0.92	0.82
	25	146	27	3	1	115	0.90	0.96	0.97	0.93
2nd derivative	5	146	39	2	2	103	0.95	0.95	0.97	0.95
	10	146	20	2	2	122	0.91	0.91	0.97	0.91
	15	146	20	4	5	117	0.83	0.80	0.94	0.82
	20	146	24	5	8	109	0.83	0.75	0.91	0.79
	25	146	26	4	0	116	0.87	1.00	0.97	0.93
SNV	5	146	35	6	1	104	0.85	0.97	0.95	0.91
	10	146	21	1	6	118	0.95	0.78	0.95	0.86
	15	146	21	3	7	115	0.88	0.75	0.93	0.81
	20	146	23	6	7	110	0.79	0.77	0.91	0.78
	25	146	25	5	0	116	0.83	1.00	0.97	0.91
Scaling	5	146	39	2	0	105	0.95	1.00	0.99	0.98
	10	146	22	0	2	122	1.00	0.92	0.99	0.96
	15	146	21	3	5	117	0.88	0.81	0.95	0.84
	20	146	23	6	6	111	0.79	0.79	0.92	0.79
	25	146	27	3	1	115	0.90	0.96	0.97	0.93

Based on the confusion matrices as shown in Figure 5, the models utilizing mean centering and 1st derivative pre-treatments exhibited the lowest classification errors, characterized by minimal FP and FN values. This demonstrates the models' robust ability to clearly distinguish between sugarcane leaf groups with different sand-mixture ratios. In contrast, models using raw spectra or min-max normalization showed higher rates of misclassification.

Figure 6 displays LDA scatter plots of components 1 and 2 further support these findings. A distinct separation between data clusters was observed in the cases of mean centering, scaling, and 1st derivative pre-treatments. Conversely, the raw spectra plot reveals significant overlapping among some sample groups. The distribution of classification scores illustrated in Figure 6 highlights a distinct structural clustering split between the lower contamination levels (5% and 10%) and the higher contamination levels (15%, 20%, and 25%) along the first linear discriminant axis (LD1). Additionally, the secondary axis (LD2) exhibits a non-monotonic progression relative to the nominal sand content increments. Because environmental variables, sample drying, and compaction pressure were kept uniform across all groups during preparation, this phenomenon is not indicative of random batch anomalies, moisture gradients, or repeated-scan artifacts.

Furthermore, because the sugarcane leaf biomass is inherently hygroscopic and was not oven-dried like the sand, minor variations in moisture content—indicated by the dominant O-H absorption band around 1450 nm—cannot be entirely ruled out as a contributing factor to the separation shown in Figure 6. Therefore, the observed transition is best framed as a proposed mechanism driven by both physical structure and potential moisture gradients.

This behavior reveals a distinct shift in the optical and structural matrix properties within the leaf pellets. At lower contamination levels (<10%), the incident NIR light penetrates a matrix primarily composed of organic sugarcane fibers, allowing the functional chemical overtones of cellulose, hemicellulose, and lignin to dictate the spectral profile. However, once the sand content matches or exceeds 15%, the concentration of inorganic silica (SiO₂) particles reaches a prominent loading on the pellet's outer surface.

The transition observed at approximately 15% sand contamination is interpreted as a shift in the optical properties of the pellet matrix. At higher contamination levels, the presence of inorganic silica increasingly alters the NIR response, shifting it from predominantly organic chemical absorption toward enhanced light scattering and specular reflection [11-12]. Rather than a strict physical threshold, this 15% boundary serves as an interpretation of the matrix behavior under high inorganic loading.

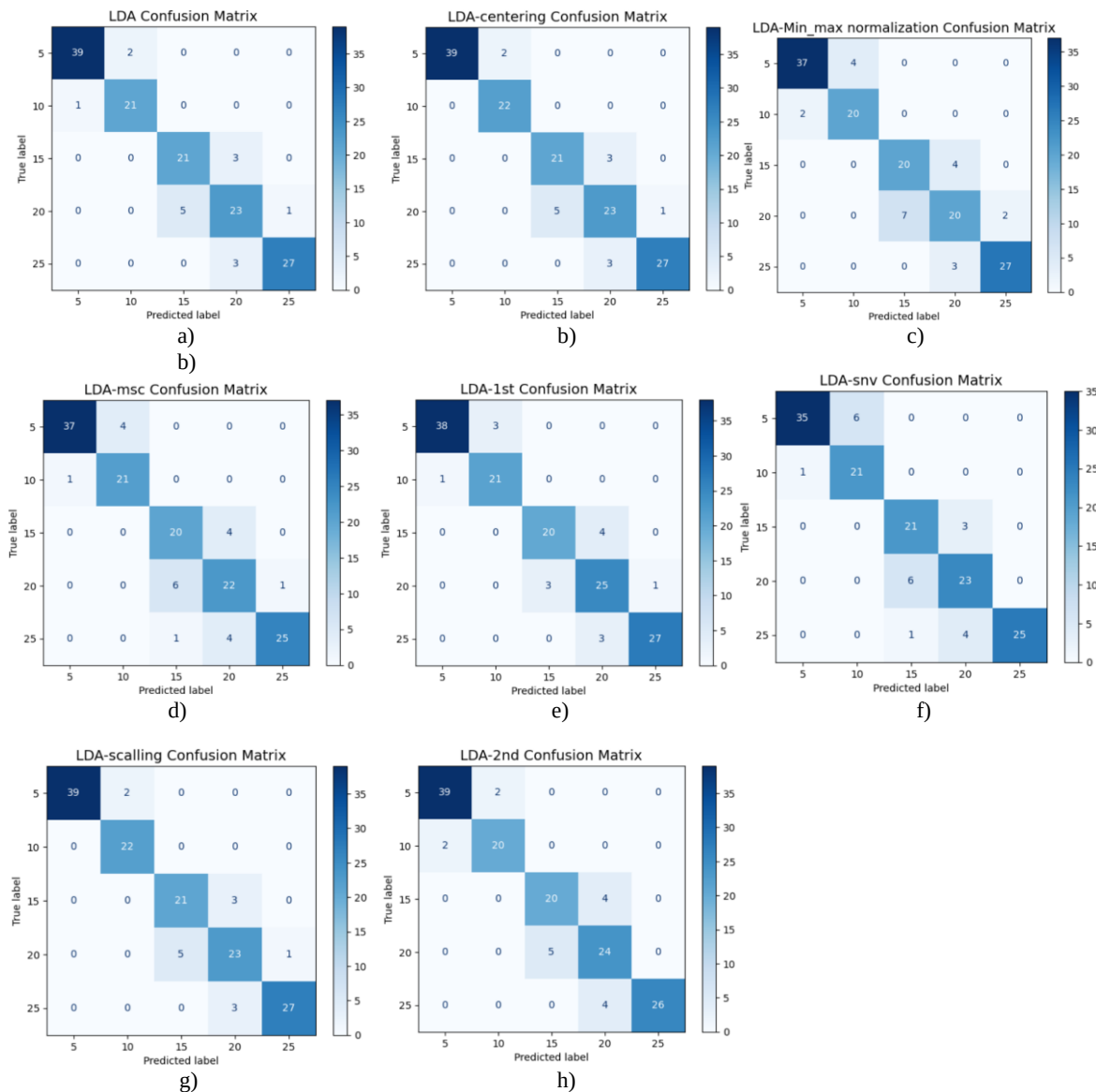


Figure 5 Confusion matrices of the LDA models developed using different spectral pre-treatments: (a) raw spectra, (b) mean centering, (c) min-max normalization, (d) MSC, (e) 1st derivative, (f) SNV, (g) scaling, and (h) baseline offset

This physical shift transitions the interaction from diffuse organic absorption to dominant scattering and specular surface reflection. LD1 successfully maps this fundamental change in optical behavior, separating the organic-dominant and inorganic-dominant regimes, while the non-monotonic path of LD2 captures the irregular, transitional surface textures and micro-density variations that occur when high ratios of sand grains compress alongside raw biomass fibers.

As the sand content ratio in the samples increased from 5% to 25%, a downward trend in the model's accuracy was observed. Specifically, in samples with a sand ratio exceeding 20%, the accuracy decreased to approximately 0.90–0.92. This decline may be attributed to the spectral interference from the sand, which potentially masked the key spectral features of the biomass.

Therefore, spectral pre-processing significantly impacts the performance of LDA models in classifying sugarcane leaf pellets with varying sand contents. The mean centering and 1st derivative methods provided the highest accuracy and the clearest separation between sample groups. These results demonstrate the high potential of NIR spectroscopy combined with LDA as a rapid, non-destructive tool for biomass fuel quality analysis.

The experimental results demonstrate that spectral pre-processing techniques significantly influence the classification performance of LDA for sugarcane leaf pellets with varying sand-mixture ratios. This was consistent with the fundamental principles of NIR spectroscopy, where analytical accuracy heavily depends on the mitigation of interference, such as light scattering and baseline shifts caused by variations in sample surface texture or density. In this study, mean centering and 1st derivative pre-treatments provided superior results by effectively eliminating noise and highlighting spectral features correlated with the chemical composition of the leaves. Specifically, the signal around 1450 nm, associated with the O–H bonds of water, played a crucial role in distinguishing between groups. The application of derivatives enhanced the resolution of peaks related to cellulose, lignin, and hemicellulose, which were the primary structural components of biomass. The LDA classification achieved accuracy exceeding 97%, indicating that appropriate pre-processing significantly improves the model's ability to differentiate samples, particularly at low-to-moderate sand ratios where biomass signals remain dominant. However, when the sand content exceeded 20%, a slight decrease in accuracy was observed. This decline is likely due to the high scattering interference from silica (SiO₂) particles, which partially masked the spectral characteristics of the sugarcane leaves.

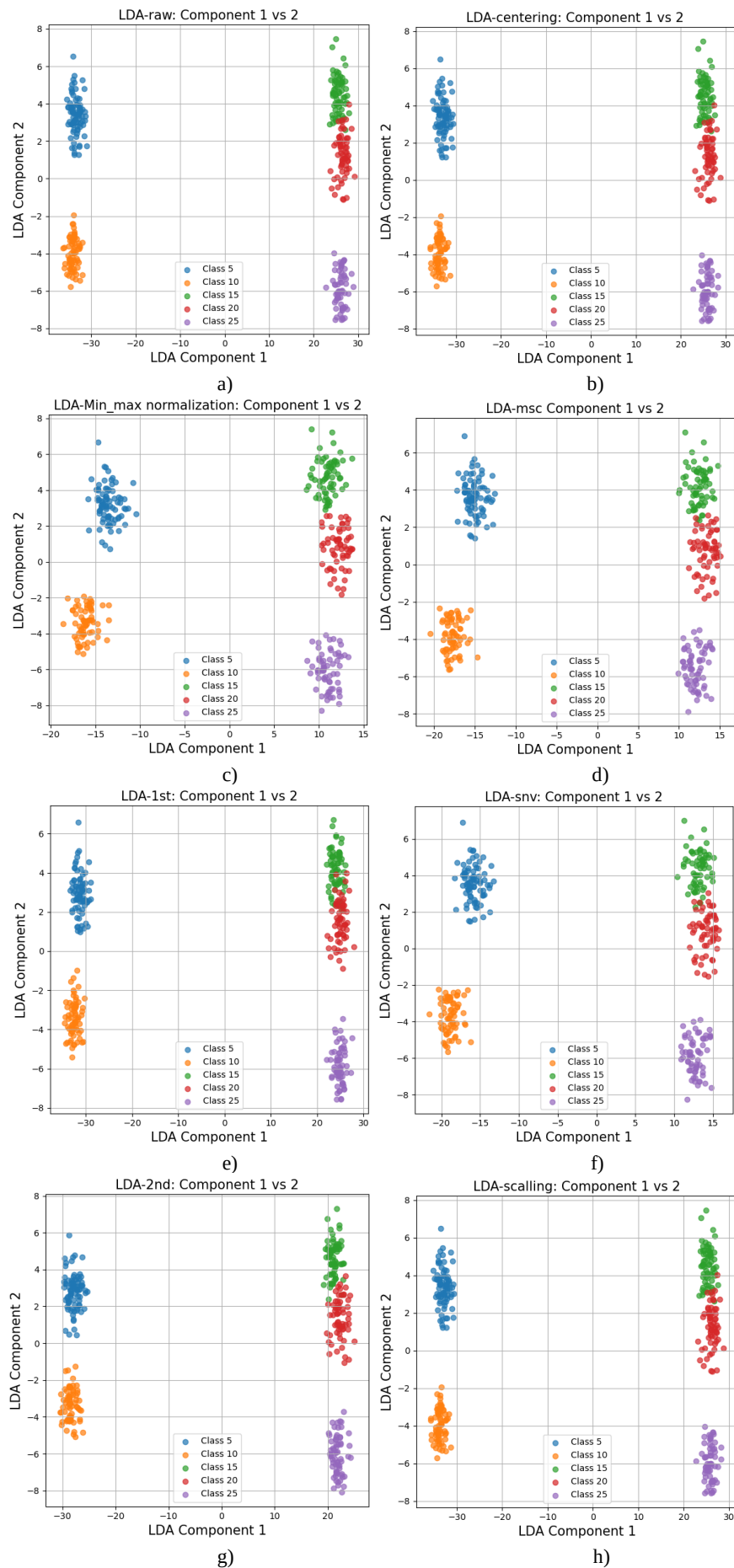


Figure 6 Scatter plot between LDA component 1 and 2 for a) Raw spectra, b) centering spectra, c) min_max normalization, d) MSC spectra, e) 1st derivative, f) SNV spectra, g) Scaling spectra and h) baseline offset

The high classification accuracy (97%) achieved by the LDA model combined with Mean Centering and first derivative preprocessing underscores the capability of NIR spectroscopy to discern complex, heterogeneous agricultural matrices. Rather than merely isolating absolute intensity variations, the combination of Centering and the First Derivative effectively removes physical light-scattering baselines caused by differences in pellet surface textures and density irregularities. This alignment allows the model to capture subtle shifts in the absorption bands associated with primary biomass constituents. As established by Feng et al. [5] in their work on two-dimensional near-infrared spectroscopic imaging for biofuel pellets, baseline scattering shifts often obscure critical chemical data. By mitigating these interferences, our model successfully amplified the sharp vibrational overtones of functional groups in sugarcane fibers, such as the C-H overtones of cellulose and lignin and the O-H bands at 1450 nm, which are highly sensitive to matrix changes. This aligns with the findings of Xue et al. [7], who noted that precise mathematical preprocessing is mandatory when isolating organic proximates in corn stover due to high overlapping signals. Furthermore, the drop in model precision observed when sand levels exceeded 20% can be interpreted through physical and optical scattering phenomena rather than simple experimental error. Sand, primarily consisting of silica (SiO_2), lacks defined, sharp absorption bands in the 860–1760 nm region but introduces heavy specular reflection and diffuse scattering. When the sand content surpasses a 20% threshold, the high concentration of inorganic particles physically dominates the sample surface. This presence shortens or severely alters the optical path length of the incident NIR light hitting the actual organic sugarcane fibers. Consequently, the signal-to-noise ratio of the essential organic overtones deteriorates, leading to the observed misclassifications in high-impurity boundaries. This threshold effect expands upon the characterization models proposed by Posom et al. [6], demonstrating that while NIR spectroscopy is exceptionally robust for predicting native organic properties and pyrolysis characteristics, an overwhelming inorganic phase introduces mechanical path-length limitations that must be accounted for in industrial, online monitoring pipelines.

The confusion matrix analysis confirmed this trend, showing that mean centering and 1st derivative models had the lowest misclassification rates, whereas raw spectra and Min-Max normalization exhibited higher error rates. Furthermore, the LDA scatter plots revealed distinct cluster separation under optimal pre-processing, reflecting the model's efficacy in extracting discriminant features from the spectral data. Technically, LDA was well-suited for datasets with relatively linear distributions, as it effectively establishes linear decision boundaries between classes. Nevertheless, to enhance accuracy in highly complex scenarios such as extreme sand concentrations or significant moisture fluctuations advanced machine learning techniques such as support vector machines (SVM), partial least squares discriminant analysis (PLS-DA), or artificial neural networks (ANN) should be considered for their superior ability to capture non-linear relationships. In terms of practical application, these findings highlight the potential of NIR spectroscopy combined with LDA as a rapid and non-destructive tool for monitoring impurities in biomass. This technology could be implemented in biomass fuel production facilities to screen raw materials for contaminants, such as sand, before the combustion process. Such screening helps reduce machinery wear and ensures consistent product quality. Additionally, the system could be integrated into production lines for on-line, real-time quality control, reducing both the time and costs associated with traditional chemical analysis. Future developments, including expanding the dataset, optimizing wavelength ranges, or employing more flexible machine learning algorithms, will further enhance precision and broaden the industrial scope of NIR analytical systems for the sustainable biomass energy sector.

4. Conclusions

This research successfully demonstrates the feasibility of using NIR spectroscopy combined with LDA to classify sand contamination levels in sugarcane leaf pellets. The study concludes that the performance of the classification model is highly dependent on the choice of spectral pre-processing. Specifically, Mean Centering and 1st Derivative techniques proved to be the most effective, as they significantly reduced light-scattering interference and baseline shifts while enhancing the spectral signatures of biomass constituents such as cellulose, lignin, and hemicellulose. The optimized LDA model achieved a remarkable classification accuracy of over 97%. Although the model performed exceptionally well across most levels, a slight decline in accuracy was noted when sand content exceeded 20%, primarily due to the reflective interference of silica (SiO_2). Despite this limitation at high contamination levels, the overall results confirm that NIR spectroscopy is a robust, rapid, and non-destructive method for impurity screening. In a practical context, this technology offers a high viability for on-line monitoring systems within the biomass energy industry. By implementing this approach, processing plants can effectively screen raw materials for sand and other contaminants before combustion, thereby preventing machinery wear, reducing maintenance costs, and ensuring consistent fuel quality. This study demonstrates the feasibility of utilizing NIR spectroscopy combined with LDA to classify discrete sand contamination levels in sugarcane leaf pellets at a laboratory scale. This non-destructive framework provides a vital benchmark for developing future automated industrial quality control systems.

5. Limitations

Although the combination of Mean Centering, First Derivative preprocessing, and LDA demonstrated high classification performance, several limitations of this study must be addressed. First, while a high volume of NIR spectra was captured, the model relied on a relatively small number of distinct physical pellets, which increases the risk of overfitting to the specific material profiles processed. Second, because the samples were prepared and evaluated under strictly controlled laboratory conditions, they do not account for external variables typical of raw industrial processing, such as ambient humidity fluctuations, variable moisture absorption, and inconsistencies in physical pellet density. Third, sand was utilized as the exclusive model impurity agent; real-world agricultural biomass often carries a heterogeneous mix of fine clay, loam, organic soil matter, and metal debris, which exhibit different light-scattering properties. Finally, model validation was restricted to an internal sample-level split. Developing a fully independent external validation protocol using separate batches sourced across different harvest seasons will be crucial to definitively confirming the robustness of the system before deploying it in commercial biomass quality monitoring.

6. Future directions

While the developed LDA model exhibits high classification accuracy under controlled conditions, it is important to note that this study was conducted on a laboratory scale utilizing prepared, homogeneous sand-and-biomass mixtures. In real-world industrial environments, factors such as fluctuating material moisture, varying particle size distribution, matrix density changes, and conveyor belt vibrations introduce significant dynamic noise. Therefore, the practical implementation of this method for automated, factory-level

online screening should be viewed as a vital direction for future research. Subsequent work will focus on validating these calibration frameworks using raw, heterogeneous industrial factory samples under dynamic sorting environments.

7. Acknowledgements

We would also like to thank the Department of Research and Graduate Studies, Khon Kaen University, Thailand, for their research support. During the preparation of this work the authors used ChatGPT 4.0 in order to check and correct grammatical errors. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

8. References

- [1] Department of Alternative Energy Development and Efficiency. Thailand's energy situation report [Internet]. 2022 [cited 20 Jan 2026]. Available from: <https://pei.dede.go.th/dataset/energy-situation/resource/e67e2a97-4d90-4edb-af24-a5d069862d4a>.
- [2] National Statistical Office. Explore basic agricultural data at the village level [Internet]. 2021 [cited 20 Jan 2026]. Available from: <https://www.nso.go.th/nsoweb/index#gsc.tab=0>.
- [3] Office of The Cane and Sugar Board. Report on the sugarcane production situation, 2024/2025 [Internet]. 2025 [cited 20 Jan 2026]. Available from: <https://www.ocsb.go.th/>.
- [4] Office of the Permanent Secretary, Ministry of Higher Education, Science, Research and Innovation. Annual Report 2018, Ministry of Science and Technology [Internet]. 2018 [cited 20 Jan 2026]. Available from: <https://www.ops.go.th/th/data-store/94-annual-report/741-mostannualreport2018>.
- [5] Feng X, Yu C, Shu Z, Liu X, Yan W, Zheng Q, Sheng K, He Yong. Rapid and non-destructive measurement of biofuel pellet quality indices based on two-dimensional near infrared spectroscopic imaging. *Fuel*. 2018;228:197-205.
- [6] Posom J, Saechua W, Sirisomboon P. Evaluation of pyrolysis characteristics of milled bamboo using near-infrared spectroscopy. *Renewable Energy*. 2017;103:653-65.
- [7] Xue J, Yang Z, Han L, Yuchen Liu, Yao Liu, Chengcheng Zhou. On-line measurement of proximates and lignocellulose components of corn stover using NIRS. *Applied Energy*. 2015;137:18-25.
- [8] Colette C. Fagan, Colm D. Everard, Kevin McDonnell. Prediction of moisture, calorific value, ash and carbon content of two dedicated bioenergy crops using near-infrared spectroscopy. *Bioresource Technology*. 2011;102(8):5200-6.
- [9] Gillespie GD, Everard CD, McDonnell KP. Prediction of biomass pellet quality indices using near infrared spectroscopy. *Energy*. 2015;80:582-8.
- [10] Posom J, Sirisomboon P. Evaluation of lower heating value and elemental composition of bamboo using near infrared spectroscopy. *Energy*. 2017;121:147-58.
- [11] Rinnan Å, van den Berg F, Engelsen SB. Review of the most common pre-processing techniques for near-infrared spectra. *Trends in Analytical Chemistry*. 2009;28(10),1201-22.
- [12] Stenberg B, Viscarra Rossel RA, Mouazen AM, Wetterlind J. Visible and near infrared spectroscopy in soil science. *Advances in Agronomy*. 2010;107,163-215.