International Journal of Design & Nature and Ecodynamics
Vol. 20, No. 11, November, 2025, pp. 2675-2686
Journal homepage: http://iieta.org/journals/ijdne

# NIR and Machine Learning-Based Rapid Monitoring of pH and Moisture in Citronella Residue Fermentation

Indra Wahyudi[1,2,3], Agus Arip Munawar[3,4], Efstathios Kaloudis[5], Sitti Wajizah[2,3], Samadi[2,3*]

[1] Doctoral Program of Agricultural Science, Postgraduate School, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia
[2] Department of Animal Science, Faculty of Agriculture, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia
[3] Research Center for Innovation and Feed Technology, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia
[4] Department of Agricultural Engineering, Faculty of Agriculture, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia
[5] Computer Simulation, Genomics and Data Analysis Laboratory, Department of Food Science and Nutrition, School of the Environment, University of the Aegean, Myrina 81400, Greece

Corresponding Author Email: samadi177@usk.ac.id

https://doi.org/10.18280/ijdne.201119

**Received:** 15 September 2025
**Revised:** 14 November 2025
**Accepted:** 22 November 2025
**Available online:** 30 November 2025

***Keywords:***
*citronella residues, feed, machine learning, NIR, ruminant, solid-state fermentation*

## ABSTRACT

The moisture content and pH level are the primary parameters influencing the efficacy of the solid-state fermentation (SSF) process of feed ingredients derived from agricultural residues. Due to the potential for contamination and process failure, the traditional methods for detecting changes in moisture content and pH level during the SSF process are unfeasible. Consequently, there is an urgent necessity to develop alternative techniques that yield highly accurate results without being time-consuming or labor-intensive. One of the most promising sensing techniques for in-line applications is near-infrared (NIR) spectroscopy. This study employed both classical and advanced machine learning (ML) models based on NIR spectra to develop a predictive model for moisture content and pH level in the thermophilic SSF process of citronella residues (CR) feed for ruminant livestock using different white-rot fungi. Principal component analysis (PCA) was utilized on the NIR spectra to extract relevant features for input into the ML models. Among the models evaluated, support vector regression (SVR) demonstrated the highest predictive accuracy ($R^2_p$ of 1.00 for both moisture content and pH level), outperforming light gradient-boosting machine (LightGBM) and random forest (RF). Although SVR achieved the highest predictive accuracy, LightGBM offers practical advantages, including faster training, lower computational demand, and better scalability for large datasets. With competitive predictive performance ($R^2_p$ of 0.95 for moisture and 0.87 for pH), LightGBM provides a strong alternative for applications requiring real-time or resource-efficient deployment. In conclusion, integrating NIR spectroscopy with ML offers a promising pathway for intelligent and real-time monitoring in large-scale SSF applications, contributing to sustainable valorization of agricultural residues into high-quality ruminant feed.

## 1. INTRODUCTION

The biotransformation of lignocellulosic biomass derived from agricultural residues represents a viable renewable resource for ruminant feed, enhancing sustainable animal production, feed security, and environmental sustainability. This approach effectively addresses the negative consequences of the inherent conflict between human and livestock food production levels, thereby posing a direct risk to food security. Citronella residues (CR), a significant by-product of the citronella distillation process, are regarded as a promising option for this purpose due to their considerable potential. Each 1,000 kg of distilled citronella leaves yields 8 kg of essential oil, while the remaining 992 kg of biomass residue is discarded as waste [1]. The CR contains 5.82% crude protein, 2.79% crude fat, and 35.03% crude fiber [2], which is composed of cellulose (35 to 40%), hemicellulose (25 to 30%), and lignin (15 to 20%) [3]. Citronella residues have been utilized as ruminant feed, though not extensively. Its extensive use in ruminant feed faces a notable challenge, specifically due to its low feed intake and digestibility [4]. The complex and extensively lignified structure of lignocellulose cell walls renders them less accessible to microbial enzymes in the rumen. Therefore, adequate pretreatment is essential before further utilization.

Various methods have been employed to convert highly lignified biomass into more digestible animal feed. Several processing methods have been employed by researchers, encompassing physical, chemical, and biological procedures. Currently, biological processing, particularly fungal microbial fermentation, has attracted considerable attention from researchers due to the increasing global demand for environmentally sustainable technologies [5]. Solid-state fermentation (SSF) processes have been used to promote the microbial fermentation of fungi on solid substrates, including citronella residues. SSF is a fermentation process that utilizes a solid substrate to support the growth of microorganisms, mimicking the conditions present in natural habitats where microorganisms grow in the absence or near absence of free water [6, 7]. Among the fungal strains used in SSF, *Phanerochaete chrysosporium*, *Pleurotus ostreatus*, *Trichoderma viride*, and *Lentinula edodes* have been studied for their strong lignocellulolytic enzyme activity, enabling effective degradation of complex fibrous substrates [5, 8-11]. SSF is characterized by lower costs, mild reactions, and an effortless manufacturing process. At present, SSF is conducted on a commercial basis within the food sector and waste treatment. Nevertheless, conventional SSF in the food and feed sector predominantly occurs under mesophilic settings (from 20 to 40°C), accompanied by an essential autoclaving procedure before fermentation to prevent product contamination.

SSF of highly lignified biomass into value-added feed faces many challenges, including the types of substrates and fungal strains, which lead to varying final product quality. The fermentation process is affected by several parameters, including pH, temperature, nutrient availability, substrate moisture content, incubation duration, and inoculation volume, which vary among different microbial species [12]. Consequently, awareness of the microbe's growth circumstances is critical for maximizing metabolite production via SSF. Among them, pH and moisture content are considered the most important process parameters for microorganism growth, cellulase production, and microbial protein synthesis during fermentation [13]. Previous studies have reported that optimal initial moisture and pH levels vary depending on the fungal strain and substrate, with most SSF protocols starting within a slightly acidic range (pH 4.5–6.0) to support fungal growth [5, 14]. In our study, initial pH was measured but not controlled throughout fermentation, as pH was expected to change naturally due to microbial activity. This approach is consistent with previous SSF studies aiming to monitor fermentation dynamics rather than maintain constant physicochemical conditions. Microbial activity affects the pH level and moisture content of the substrate, hence influencing the quality of the fermentation output. The main reason for changes in pH during SSF is the release of organic acids, such as lactic, citric, and acetic acids. On the other hand, the increase in pH is linked to the assimilation of these organic acids [12]. Meanwhile, changes in moisture content in the substrate play an important role in microbial growth, enzyme production, and nutrient transfer [15]. Therefore, monitoring pH level and moisture content during the SSF process is crucial to achieve an optimum quality and yield of the desired product. To achieve this goal, an analytical method is required that can deliver real-time data on critical process parameters. These parameters are usually measured offline, which takes time and adds to analytical error through sampling and sample preparation [16].

One of the most promising sensing techniques for in-line applications is near-infrared (NIR) spectroscopy. NIR characterizes materials by assessing the absorption or reflection of light at wavelengths ranging from 850 to 2,500 nm. Owing to the inherent variations in the chemical composition and physical quality of materials, different spectra can be generated for each material. The NIR method serves as an effective and reliable tool for real-time monitoring of bioprocess and shows potential for future applications in intelligent control of feed production. This approach has been utilized to create rapid, accurate, non-destructive, and reproducible techniques for analyzing the compositions of diverse materials, such as food [17], agricultural products [18, 19], quality of fermentation products [20], animal feed [21, 22], and so on. The NIR spectral range (1,000–2,500 nm) has been widely used for the evaluation of fermented lignocellulosic biomass due to its ability to capture molecular vibrations associated with organic components such as moisture, proteins, and fiber. For instance, Dai et al. [23] successfully applied NIR spectroscopy for the rapid and cost-effective determination of pH, moisture, soluble protein, and trypsin inhibitor contents during thermophilic SSF of unsterilized soybean meal by *Bacillus licheniformis* YYC4. From an application perspective, NIR can quickly measure the moisture content of substrates, help maintain microbial activity, and determine the appropriate decision-making method. NIR enables rapid assessment of pH level, which is essential for assessing microbial metabolic activity to ensure optimal product results [12]. However, the complexities of NIR spectra present several challenges for data interpretation. Numerous wavelengths and absorption bands in the NIR spectra complicate the analysis. This complexity leads to too much interference and multicollinearity, reducing prediction stability. In SSF systems, variations in particle size and density inhomogeneity further affect light scattering and spectral consistency. Such effects can be detrimental in real-time applications where reliability and consistency of predictions are essential [24]. Consequently, the ability to generalize predictions on new data sets often shows poor model prediction accuracy in real-time application scenarios. In recent years, improvements in computational machine learning (ML) algorithms have made it a more useful tool for data mining and building models. Studies indicate that ML algorithms demonstrate superior performance in both qualitative and quantitative predictions of materials when employing NIR, Hyperspectral, and Raman spectroscopy [25, 26]. This finding provides insights and opportunities for integrating NIR technology with ML algorithms to predict pH level and moisture content during the SSF process. However, it is still uncertain which ML algorithm is capable of improving the accuracy of the SSF monitoring process using NIR technology, particularly in the context of animal feed bioprocesses. The objective of this work is to utilize various ML algorithms in conjunction with features extracted from NIR spectra using principal component analyses (PCA) to develop highly accurate predictive models for pH level and moisture content. As a result, this model offers an innovative, efficient, and high-throughput approach for monitoring animal feed bioprocess and provides rapid technical support for precise decision-making in large-scale SSF systems.

## 2. MATERIALS AND METHODS

### 2.1 Substrate and fungal strains

Citronella residues (CR) served as the substrate for sequential fermentation in this study. CR was collected from farmers in the Gayo Lues District of Aceh, Indonesia, after the extraction of essential oil. The wet residues of CR were carefully processed to achieve an average particle length of approximately 3 cm. This particle size was chosen based on preliminary trials to optimize fungus accessibility and aeration during fermentation. The chopped material was oven-dried at

60°C for roughly 48 hours until reaching a final moisture content of 10–12%, preserving the structural carbohydrates (cellulose and hemicellulose) without thermal degradation [27]. The dried substrate was then stored in airtight containers until further fermentation processing. For fermentation microbes, this study utilized various fungal strains, including *P. chrysosporium* (PCH), *P. ostreatus* (POS), *T. viride* (TRV), and *L. edodes* (LED). The four fungal strains selected for this study were chosen based on their reported ligninolytic and cellulolytic enzyme activities and prior successful use in lignocellulosic biomass fermentation [5, 14]. These strains are known to improve fiber degradation and nutrient availability in various agricultural by-products, making them suitable candidates for SSF of citronella residues. The microbial strains were sourced from the Indonesian Culture Collection (InaCC) Laboratory of BRIN, Cibinong, Indonesia. Before fermenting CR, the fungal strains were pre-cultivated following the protocol of Tuyen et al. [14] with minor modifications. Specifically, the fungus was grown on Potato Dextrose Agar (PDA) medium and incubated at 24°C until its mycelia had extensively colonized the agar surface. Inoculum preparation involved transferring an agar fragment (from 1.5 to 2.0 cm) containing fungal culture onto sterilized cracked corn. The inoculated corn was then incubated at 24°C until it was fully colonized by fungal mycelia. To maintain the inoculum and inhibit further growth, the corn grain spawn was stored at 6°C in a controlled environment for one week prior to fermentation.

## 2.2 Fermentation of CR samples over time periods

In this study, SSF was initiated using 447 g of dried CR as the solid substrate. The substrate matrix was further enriched with nutrients, including 30 g of molasses and 100 g of corn bran. Subsequently, the substrate was inoculated with 50 g of corn grain spawn of each fungal strain (PCH, POS, TRV, and LED), and sterile water was added during mixing to maintain a total moisture content of 60%. Both the SSF and uninoculated substrate (WOI) were aerobically incubated at room temperature (approximately 37°C) for 28 days in polyethylene bags. Four full fermentation periods were carried out in five replicates, and each period lasted seven days with 25 data sets. Samples used for NIR spectra acquisition and laboratory reference measurements were taken every seven days during the fermentation process. Thus, a total of 100 samples were obtained in four different SSF periods.

## 2.3 NIR spectra acquisition

The NIR spectra of the SSF samples were collected using the NIRFlex N-500 spectrometers (Büchi, Flawil, Switzerland). Spectra measurements were performed on the SSF samples immediately after harvesting during each period. About 5 g of each SSF sample was placed in a sample holder and flattened to create a smooth surface before scanning. NIR spectra were measured in the absorbance mode in the wavelength range from 1,000 to 2,500 nm (10,000 to 4,000 cm$^{-1}$) with an average resolution of 1 nm, resulting in 1,557 data points. Each spectrum was scanned 32 times, and the results are averaged per single spectrum. Spectra acquisition was conducted at an ambient temperature of approximately 29 to 31°C.

## 2.4 Measurements of moisture content and pH level

The moisture content of the sample was determined based on the AOAC 930.15 method by drying 2 g of SSF product to constant weight at a temperature of 103 ± 2°C. The moisture content was determined by calculating the weight loss following the drying process, and it is expressed as a percentage on the wet basis. The pH was measured with a pH meter by weighing 1 g of the sample and mixing it well with 50 mL of deionized water. The solution was centrifuged at 4,000 revolutions per minute for 10 minutes. Following centrifugation, pH was assessed.

## 2.5 Data partitioning and PCA dimension reduction

All samples were partitioned into calibration and prediction sets at a 4:1 ratio, comprising 80 and 20 samples, respectively. The calibration set was used for model development, while the prediction set served for independent validation of predictive performance. To ensure balanced representation, data were stratified according to fermentation time and fungal strains. Specifically, within each fungal treatment and each fermentation phase, one out of every five samples was systematically allocated to the prediction set, while the remaining four were retained in the calibration set. This procedure maintained proportional representation across all fermentation periods (7, 14, 21, and 28 days) and fungal strains (WOI, PCH, POS, TRV, and LED). Additionally, samples exhibiting the highest and lowest pH and moisture values were included in the calibration set to capture the full range of spectral variability and enhance model robustness. This stratification strategy ensured that both data subsets reflected the overall spectral diversity of the experiment, supporting reliable model evaluation and generalization.

Principal component analysis (PCA) was applied to the preprocessed NIR spectral data (1,000–2,500 nm) to reduce data complexity and remove redundancy among the spectral variables [28]. The data matrix was structured with samples as rows and absorbance values at each wavelength as columns. Prior to analysis, the data were mean-centered and autoscaled to ensure that all wavelengths contributed equally. PCA was performed using the scikit-learn library in Python, and the number of principal components (PCs) retained was determined based on the criterion of eigenvalues greater than 1 and the cumulative proportion of explained variance [29]. A total of 10 principal components were retained, accounting for over 90% of the total spectral variation. The resulting PCs were used as input variables in the development of predictive models for estimating moisture content and pH.

## 2.6 Machine learning (ML) models

The ML modeling workflow consisted of three primary steps: (1) defining the algorithm structure, (2) tuning hyperparameters, and (3) evaluating performance. In this study, three different types of ML algorithms were employed to develop models for predicting moisture content and pH, and their performances were compared. These algorithms covered both classical and advanced ML methodologies. Initially, random forest (RF) and support vector regression (SVR) were used to establish classical ML models. For the RF model, structural parameters included the number of decision trees and maximum tree depth. The SVR model was characterized by the kernel type (radial basis function), regularization parameter (C), and kernel coefficient (γ). Additionally, light gradient-boosting machine (LightGBM), an advanced gradient-boosting machine (GBM) algorithm, was selected as the primary algorithm to optimize the GBM framework. LightGBM was chosen due to its leaf-wise tree growth strategy and its ability to discretize continuous values into bins, which significantly enhances training speed and memory efficiency [30]. For LightGBM, the structure was defined by the number of leaves, maximum depth, and boosting iterations.

Hyperparameters such as learning rate, maximum tree depth, minimum child weight, and the number of boosting iterations were optimized using Bayesian optimization during model calibration. Bayesian optimization helps prevent overfitting by penalizing overly complex solutions and promoting models that generalize well to unseen data. This optimization provides an efficient way to explore the hyperparameter space by iteratively assessing model performance, typically measured by root mean square error (RMSE), and updates a probability model to identify the most promising hyperparameter configurations [31]. Subsequently, k-fold cross-validation was applied during model calibration by dividing the data into 10 folds, with each fold alternately serving as a validation set while the others functioned as the training set to balance bias and variance [32]. This process was repeated until each fold was validated once. The optimized models were then tested on independent samples to evaluate their predictive performance for moisture content and pH level.

### 2.7 Performance evaluation of prediction models

The performance of prediction models is assessed based on the squared correlation coefficient ($R^2_c$ and $R^2_p$) and the root mean square error ($RMSE_c$ and $RMSE_p$) indexes in the calibration and prediction sets. An $R^2$ value > 0.8 indicates a model with strong predictive ability, while an RMSE lower than the actual standard deviation (SD) indicates superior predictive performance [22, 24]. These metrics were calculated as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2} \quad (1)$$

where, $y_i$ is the observed value, $\widehat{y_i}$ is the predicted value, and $n$ is the number of samples.

$$SD = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (2)$$

where, $\bar{y}$ is the mean of observed values.

In addition, residual predictive deviation (RPD) and range error ratio (RER) indexes serve as further indicators. RPD is determined by dividing the actual SD by $RMSE_p$, where an RPD value of > 3 indicates a model with excellent predictive performance [29]. Meanwhile, RER is obtained from the ratio of the data range to RMSE, with an RER value of > 10 confirming a high prediction model to accurately quality control in new samples [33]. This evaluation ensures the statistical validity of the model and its application in real conditions.

### 2.8 SHAP analysis

The Shapley additive explanations (SHAP) values are used to assess the significance of features in the optimal ML model. SHAP, based on game theory, improves the interpretability of tree models by combining the local contribution of each feature for global analysis [34]. This method breaks down the prediction into the contribution of each feature, which is expressed in a positive or negative importance value, reflecting the direction of its influence. This approach allows for the identification of the most influential features and their interactions in producing the final prediction, thus increasing model transparency and overcoming the 'black box' problem in ML. The SHAP value used in the study aims to interpret the importance of wavelengths in the best predictive model.

All analyses in this study were conducted on the Google Colab platform (12 GB RAM), utilizing Python version 3.1.3 to implement statistical analyses, ML models, and graphical visualizations.

## 3. RESULTS AND DISCUSSION

### 3.1 Reference content statistics

The moisture content and pH of samples collected from various fermentation periods were initially analyzed using traditional chemical methods as a reference. Descriptive statistics for reference content of moisture and pH in calibration and prediction datasets are shown in Table 1.

**Table 1.** Descriptive statistics for reference content of moisture (% wet basis) and pH level in calibration and prediction datasets

| Statistical Parameter | Moisture (% wet basis) | | pH Level | |
|---|---|---|---|---|
| | Cal | Pred | Cal | Pred |
| n | 80 | 20 | 80 | 20 |
| Range | 13.3 | 8.7 | 5.0 | 3.5 |
| Min | 52.5 | 53.6 | 5.0 | 5.4 |
| Max | 65.8 | 62.3 | 10.0 | 8.9 |
| Mean | 58.3 | 58.2 | 7.0 | 6.9 |
| SD | 2.7 | 1.8 | 1.3 | 1.0 |

Cal: calibration; Pred: prediction; SD: standard deviation; Min: minimal; Max: maximal; n: number of sample datasets.

The moisture content and pH level of the calibration set ranged from 52.5% to 65.8% and 5.0 to 10.0, respectively. The corresponding values for the prediction set samples were 53.6% to 62.3% and 5.4 to 8.9, respectively. In general, the moisture content and pH range of the calibration set covered the entire range of the prediction set. Additionally, results of the Two-Sample Kolmogorov-Smirnov Test showed no significant differences for moisture content (P > 0.39) and pH level (P > 0.96) between the two datasets, confirming that the calibration set adequately represented the variation in the prediction set. Figure 1 illustrates that the moisture content and pH of fermented CR exhibit a gradual increase with prolonged fermentation time and display a dynamic trend throughout the different treatments. This indicates that microbial growth and metabolism vary depending on the fungus type.

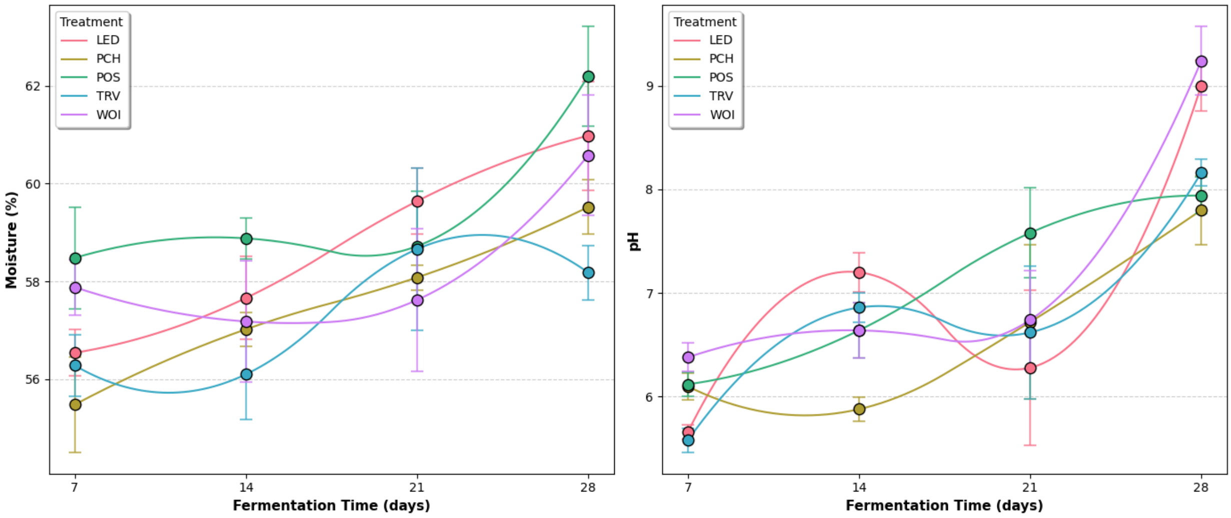


**Figure 1.** Dynamics of changes in moisture content and pH level of CR fermented with various fungal strains
LED: *L. edodes*; PCH: *P. chrysosporium*; POS: *P. ostreatus*; TRV: *T. viride*; WOI: uninoculated substrate.

In SSF-based enzyme production using agricultural waste substrates, moisture balance is a key parameter with an optimal range that varies significantly between 50 to 75%. Studies show that increasing the moisture content from 40 to 55% contributes to increased enzyme production. During vegetative growth, white-rot fungi secrete decomposing enzymes, including laccase, lignin peroxidase (LiP), and manganese peroxidase (MnP), which partially degrade complex carbohydrates into $CO_2$ and $H_2O$. This process contributes to a reduction in the overall content of neutral detergent fiber (NDF) and acid detergent fiber (ADF) in the substrate [14]. This phenomenon is related to the increase in moisture content from 52.5 to 65.8% as the duration of fermentation increases from 7 days to 28 days (Table 1), caused by the decomposition of the substrate by microbes, so that the need for enzyme production becomes lower. The observed increase in moisture content during the fermentation period is likely due to a combination of microbial metabolic activity and the breakdown of lignocellulosic components, which can release bound water into the substrate matrix. Additionally, metabolic water produced during the aerobic degradation of carbohydrates and proteins may have contributed to the overall moisture accumulation. The relatively closed incubation environment may also have limited evaporation, promoting moisture retention in the substrate. pH is another important parameter in the context of enzyme production through SSF. However, it is typically not a primary focus in SSF and is mostly maintained during the initial stages by keeping the moisture of the substrate. Changes are possible during the enzyme production process due to microbial metabolic activity. The major reason for the change in pH from 5.0 to 10.0 during SSF is due to the assimilation of organic acids such as acetic, citric, and lactic acid, resulting in an increase in pH level [17]. Filamentous fungi exhibit growth across a broad pH range of 2 to 9, with an optimal pH range between 3.8 and 6. Bacterial contamination of molds and yeasts can be diminished by adjusting the pH to levels harmful for bacterial growth [8, 35].

## 3.2 NIR spectra visualisation and PCA feature extraction

The NIR raw spectra (Figure 2) showed distinct trends in absorption among the different treatment groups (WOI, PCH, POS, TRV, and LED) throughout the fermentation period. Although all spectra exhibited similar peak positions, their relative intensities differed, indicating variations in composition among the fungal strains during fermentation.

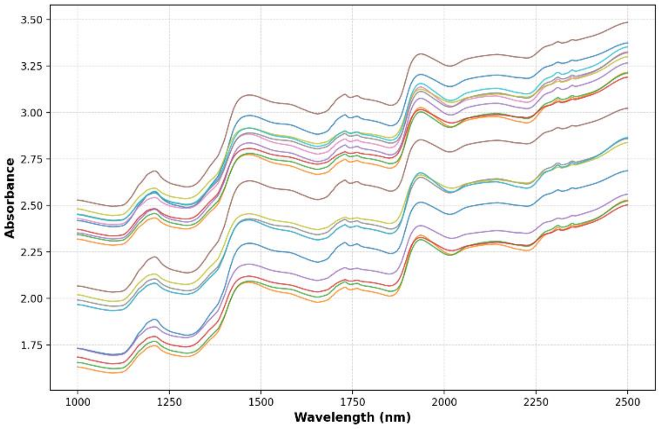


**Figure 2.** Average raw spectra of samples categorized by inoculum types, collected at fermentation times

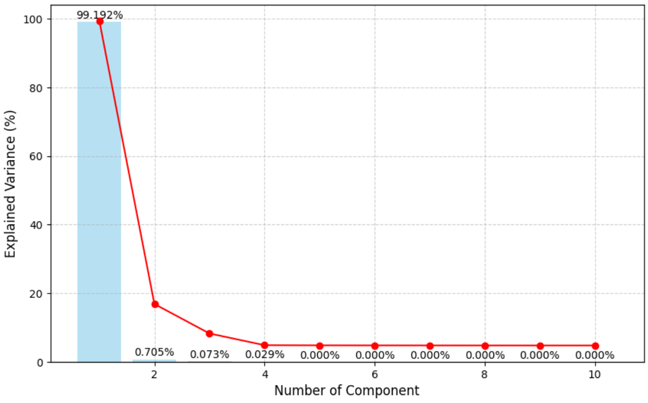


**Figure 3.** The PCA scree plot illustrates the explained variance proportion allocated to each PC

PCA was performed on raw NIR spectra to investigate natural variation, reduce dimensionality, and extract features. PCA converts NIR spectra into principal component (PC) scores, retaining the most relevant parts while eliminating collinearity, thereby enhancing ML model performance and accelerating computational processes [36]. The scree plot (Figure 3) shows that the first two PCs explain most of the total variance (PC1 = 99.19% and PC2 = 0.70%), indicating that these components capture the primary spectral differences across samples. The PCA score map (Figure 4) further illustrates the spatial distribution of samples in the PC1–PC2 space, revealing clear clustering patterns according to fermentation time and treatment. This separation confirms that NIR spectroscopy can effectively distinguish different fermentation states based on spectral characteristics. Consequently, these findings provide a foundation for directing subsequent analysis toward these components. In the ML model training process, the first 10 PCs are utilized due to the spectral variations observed in the higher-ranked components following the PCA transformation. Relationships between variables are widely investigated to determine the physical importance of PC in PCA [37].

The loading plot for the initial four PC (Figure 5) illustrates distinct absorbance patterns associated with molecular vibrations. The loading plot shows peaks and valleys that are significantly correlated with the molecular vibrations in the sample. The peaks detected at wavelengths of 1,210 nm, 1,350 nm, 1,403 nm, 1,690 nm, 1,750 nm, 1,930 nm, 2,010 nm, 2,100 nm, 2,229 nm, and 2,400 nm indicate specific molecular vibrations associated with the water content and organic compounds inside the material. Variations in moisture and pH during SSF are intricately linked to lignocellulose degradation, which can be monitored by NIR spectroscopy. The enzymatic decomposition of lignocellulose during SSF releases soluble organic acids and alters the hydrogen-bonding environment of hydroxyl (O–H) and carbonyl (C=O) groups, leading to detectable changes in overtone and combination absorption bands in NIR spectra, particularly within the 1,100-1,500 nm and 2,100-2,300 nm regions associated with pH variation. This mechanism is supported by comparable findings in thermophilic SSF of soybean meal, where NIR spectroscopy successfully monitored pH and other biochemicals [38].

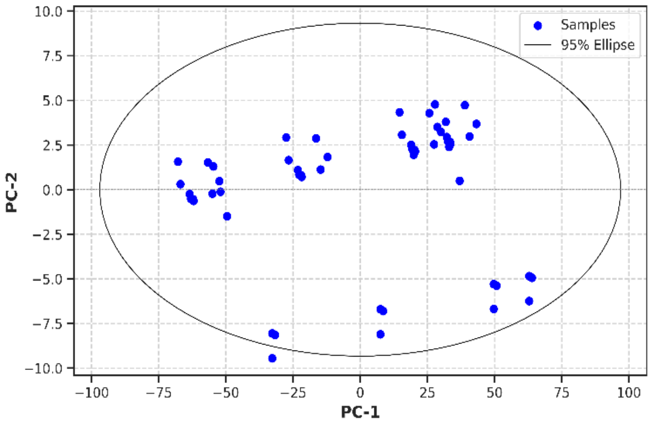


**Figure 4.** The PCA score plot based on the NIR spectral data

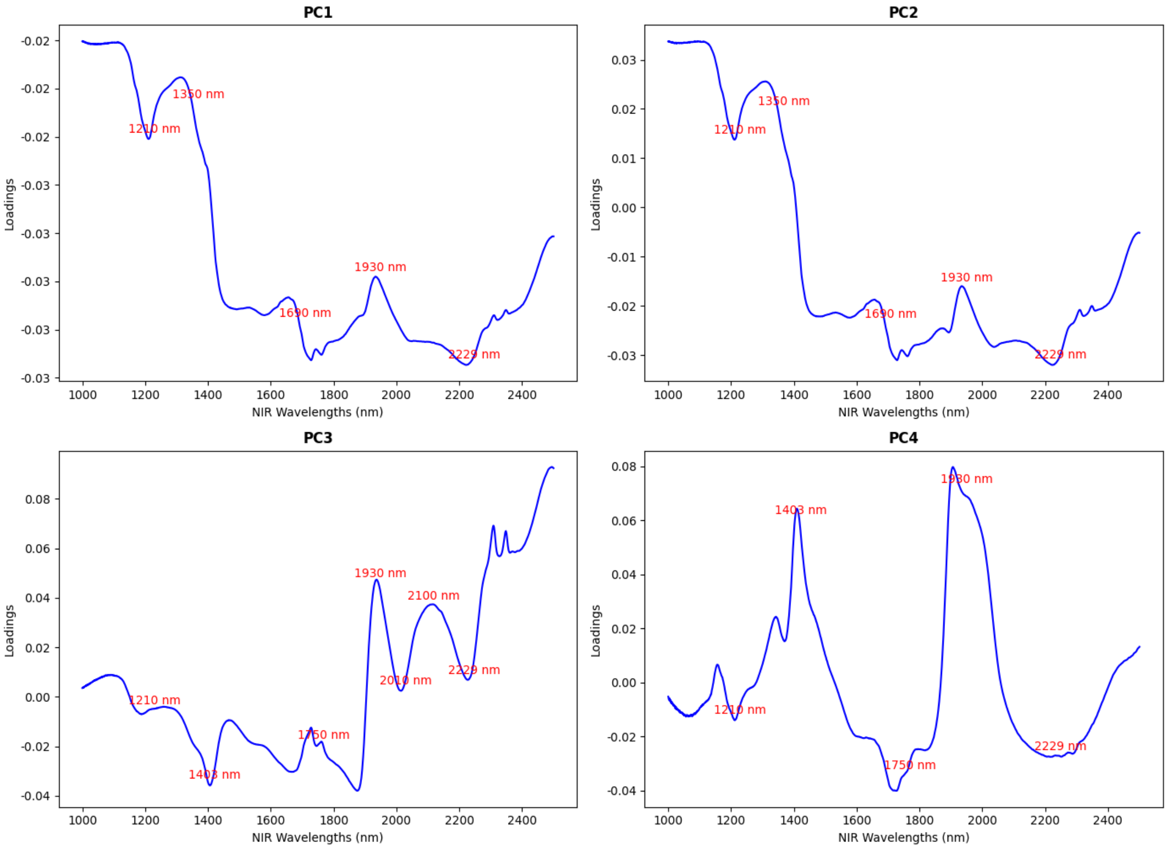


**Figure 5.** The loading plots of the four principal components (PC) from the PCA analysis of NIR spectra in the wavelength range from 1,000 to 2,500 nm

PC1: the primary principal component, which accounts for the greatest variance in the dataset. PC2: the second principal component, which accounts for the second greatest variance; PC3: the third highest variance; PC4: the fourth highest variance.

Throughout fermentation, the fungus generates lignolytic and cellulolytic enzymes that decompose lignocellulose into less complex components. Based on these results, the moisture content in the sample can be identified through intense absorbance regions at wavelengths around 1,400 nm and 1,900 nm. These bands correspond to the overtone of stretching the O–H bond in water and carbohydrate molecules [39]. In addition, the combination band that appears in the range of 2,100 nm to 2,300 nm (connected with the combination vibration of O–H, N–H, and C–H), also makes an important contribution to the characterization of moisture content [38]. This spectral region is mostly made up of information from the PC1 and PC4. These spectral regions have been consistently reported as highly correlated with sample moisture content in various agricultural and fermentation substrates. The first overtone of O–H stretching around 1400 nm and the combination band near 1,900 nm are particularly sensitive to changes in water content [39, 40]. Additionally, the 2,100–2,300 nm range, arising from O–H, N–H, and C–H combination vibrations, has been shown to contribute to moisture prediction accuracy in solid substrates undergoing bioprocessing [39]. The pH variation in the sample is correlated with changes in absorbance at wavelengths approximately between 1,100 to 1,500 nm and 2,100 to 2,300 nm, which are generally associated with overtone vibrations and combinations of functional groups such as O–H, N–H, and C=O. PC1, PC2, and PC4 exhibit significant variations within this wavelength region, thereby serving as primary indicators for inferring pH level using NIR spectroscopy. This finding facilitates the creation of ML predictive models capable of precisely quantifying moisture content and pH by leveraging spectral properties revealed through PCA transformation. This approach not only improves prediction accuracy but also avoids overfitting.

### 3.3 Performance of ML prediction model

Machine learning (ML) methods are employed to develop models for detecting moisture content and pH level, utilizing the PCA-transformed spectra dataset. The performances of the three distinct ML approaches show significant variations, as presented in Table 2.

**Table 2.** Predictive performance of RF, SVR, and LightGBM in monitoring moisture content and pH level during the SSF process of fermented CR

| Content | Method | Cal | | Pred | | | |
|---|---|---|---|---|---|---|---|
| | | $R^2_c$ | RMSE$_c$ | $R^2_p$ | RMSE$_p$ | RPD$_p$ | RER$_p$ |
| Moisture (%) | RF | 0.97 | 0.46 | 0.88 | 0.60 | 3.00 | 14.5 |
| | SVR | 1.00 | 0.03 | 1.00 | 0.04 | 45.00 | 217.5 |
| | LightGBM | 1.00 | 0.02 | 0.95 | 0.39 | 4.62 | 22.3 |
| pH level | RF | 0.93 | 0.35 | 0.77 | 0.48 | 2.08 | 7.3 |
| | SVR | 1.00 | 0.02 | 1.00 | 0.03 | 33.33 | 116.7 |
| | LightGBM | 1.00 | 0.08 | 0.87 | 0.35 | 2.86 | 10.0 |

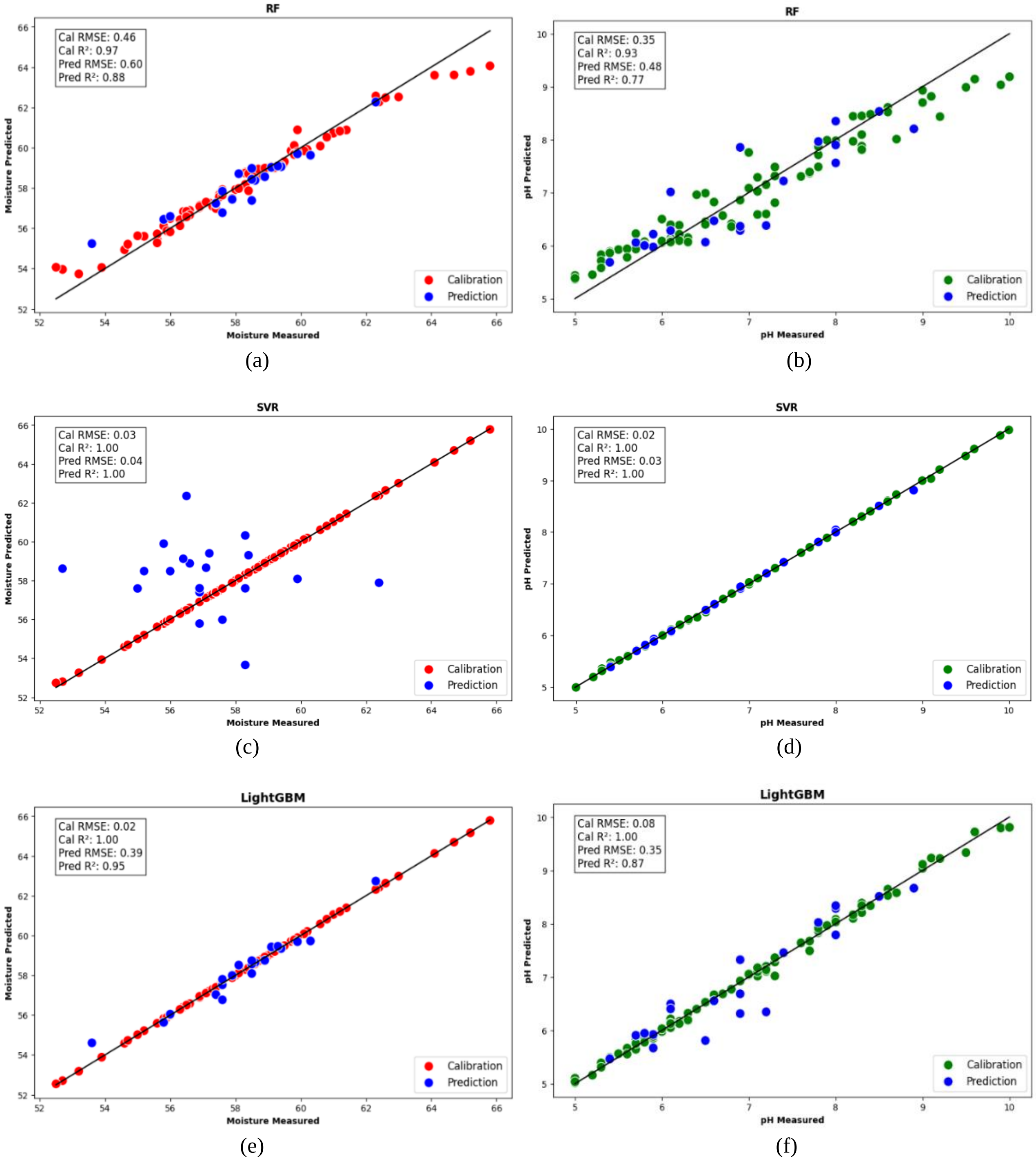
Cal: calibration; Pred: prediction; $R^2_c$: coefficient of determination for calibration; RMSE$_c$: root mean square error of calibration; $R^2_p$: coefficient of determination for prediction; RMSE$_p$: root mean square error of prediction; RPD$_p$: ratio of performance to deviation for prediction; RER$_p$: range error ratio for prediction; RF: random forest; SVR: support vector regression; LightGBM: light gradient-boosting machine.

As shown in Table 2, the performance of the ML model in predicting moisture content and pH level during the fermentation of CR ruminant feedstuffs is essential for assessing prediction accuracy. The SVR model exhibited optimal performance in predicting moisture content, achieving an $R^2$ value of 1.00 and a minimal RMSE in both the calibration and prediction phases. The model exhibited excellent generalization ability, evidenced by an RPD value exceeding three and an RER greater than 10, reflecting its high accuracy and reliability in analyzing moisture content during SSF processes.

In contrast, decision tree-based models, including RF and LightGBM, exhibited inferior performance in predicting moisture content, with $R^2$ values of 0.97 and 1.00, respectively, and a higher RMSE than SVR. An RMSE value less than the standard deviation of the target measurement indicates that the prediction error of the model is smaller than the inherent variability present in the dataset [21, 24]. Among these two models, LightGBM demonstrated superior performance compared to RF, achieving RPD values of 4.62 and an RER of 22.3, thereby categorizing it as an acceptable model for predicting moisture content. For pH prediction, a similar trend was observed, where the SVR model continued to be the best performer, with an $R^2$ value of 1.00 and a minimum RMSE during both calibration and prediction phases (RMSE$_c$ of 0.02 and RMSE$_p$ of 0.03). The LightGBM model was slightly lower than SVR but still outperformed RF, achieving an $R^2_p$ of 0.87 and RMSE$_p$ of 0.35.

However, the exceptionally high $R^2_p$ values obtained by the SVR model should be interpreted with caution. Such near-perfect accuracy is uncommon in practical bioprocess monitoring and is likely influenced by several experimental factors. First, the prediction set consisted of 20 samples, and its entire variability range was fully represented within the calibration set, which can artificially inflate apparent accuracy. Second, the fermentation was carried out under controlled laboratory conditions that minimized sample heterogeneity and reduced spectral noise. Third, PCA dimensionality reduction removed most collinearity and baseline variation, creating a smoother, more linearly separable feature space that is highly favorable for SVR, particularly when optimized with an RBF kernel through Bayesian tuning. These factors collectively contribute to the high $R^2_p$ values observed in this study but may limit generalizability to more heterogeneous, real-world SSF systems. Therefore, future studies should incorporate completely independent batches with broader variability to rigorously assess model robustness and mitigate the risk of overfitting.

The superior performance of SVR over tree-based models can also be explained by the nature of the feature space after PCA transformation. PCA extracts orthogonal components that capture the major sources of spectral variance while removing noise and nonlinear redundancy. This transformation often yields a feature space with smoother gradients and higher linear separability. SVR, particularly when using an RBF kernel, is designed to exploit such feature structures by mapping them into a high-dimensional space where an optimal separating hyperplane can be constructed. Consequently, SVR is highly sensitive to subtle but informative variations in the PCA scores. In contrast, tree-based algorithms such as RF and LightGBM partition the feature space through axis-aligned splits, making them less capable of capturing small, continuous changes within PCA-reduced data. As a result, these models may overlook the finer spectral differences that contribute to accurate prediction, explaining their comparatively lower performance in this study. The scatter plot depicting the performance of three different ML algorithms in predicting moisture content and pH level is presented in Figure 6.



(a) (b)

(c) (d)

(e) (f)

**Figure 6.** Scatter plot comparing reference and prediction values of moisture and pH level across all three machine learning models; (a) RF model for moisture content prediction, (b) RF model for pH level prediction, (c) SVR model for moisture content prediction, (d) SVR model for pH level prediction, (e) LightGBM model for moisture content prediction, and (f) LightGBM model for pH level prediction

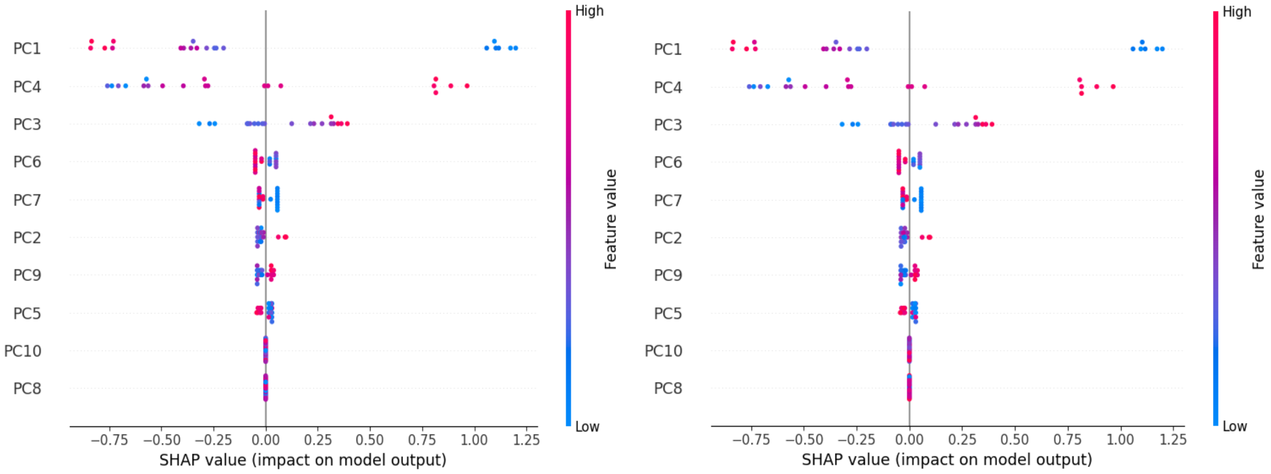
The dots represent the data points for calibration and prediction, and the black diagonal line serves as a reference for a perfect fit.

Figures 6(a)-(f) illustrate that a strong linear correlation exists between the prediction and the reference value, as evidenced by the symmetrical distribution of data points around the line. This pattern indicates that the model's predictions are reliable and free from systematic bias, thereby underscoring the model's effectiveness and robustness [41]. Figures 6(c) and 6(d) demonstrate that SVR effectively handles complex linear relationships within the reduced PCA data, leading to highly accurate predictions. PCA reduces data dimensionality and keeps the most significant PC. The application of PCA for data reduction enhances the efficacy of SVR on the processed dataset by eliminating irrelevant or redundant information. SVR is an ML algorithm for linear and non-linear data that maps data to higher dimensions for linear separation [42]. SVR searches for an optimal hyperplane to divide samples with support vectors. This creates a more favorable environment for SVR to achieve linear separation, even in cases where the initial data is complex and nonlinear. However, SVR has the disadvantage of requiring a long training period. LightGBM is also an attractive alternative, particularly when a lightweight model with sufficient accuracy is required [39].

Compared to traditional methods, this proposed approach offers a non-destructive alternative, making it a promising solution for real-time monitoring of fermentation change patterns during the SSF process, which are important for the success of CR fermentation in producing value-added products for ruminant feed. These regulations affect on growth and metabolic regulation of microorganisms, in this case, white-rot fungi. During vegetative growth, white rot fungi secrete enzymes that decompose macromolecular substances to obtain carbon and nitrogen. Tuyen et al. [14] demonstrate that white rot fungi partially degrade carbohydrates into $CO_2$ and $H_2O$, resulting in a general decrease in the NDF and ADF content of the substrate. Simultaneously, during the degradation of complex carbohydrates, the assimilation of enzymatic by-products, such as organic acids, alters hydrogen ion concentrations, leading to an increase in pH level [8]. This phenomenon explains the increase in moisture content and pH level during the fermentation process, as shown in Figure 1. Research by Pensupa et al. [43] supports this finding, indicating that spore formation on the substrate surface correlates positively with increased moisture content. This indicates that increased humidity is not only related to a higher rate of fungal growth but also supports the overall proliferation of mold colonies, which ultimately accelerates the fermentation process and improves the quality of the final product.

### 3.4 SHAP interpretable model

Beyond achieving high predictive accuracy, understanding the underlying rationale for each prediction is essential in the development of robust predictive models. The inherent complexity of machine learning models, often characterized as 'black box' systems, poses challenges in interpreting their outputs [44, 45]. To address this, SHAP values (Shapley Additive exPlanations) are utilized to quantify the contribution of specific wavelengths to the PC values within the optimal prediction model, employing the Python SHAP module. This approach highlights the importance of feature interpretability in model assessment. The impact of each principal component in the optimal SVR prediction model is visualized through the SHAP beeswarm plot in Figure 7.



(a)

(b)

**Figure 7.** SHAP beeswarm plot of the best predictive models of SVR for (a) moisture content and (b) pH level
Data points with positive SHAP values (located to the right of the vertical zero line) indicate that observations with relatively low absorbance (represented by blue shades) or high absorbance (represented by pink shades) contribute to an increase in the model output. Conversely, data points with negative SHAP values (positioned to the left of the vertical zero line) correspond to a decrease in the model output.

Based on Figures 5(a) and 5(b), the prediction results for moisture content and pH level are significantly influenced by PC1, followed by PC4 and PC3. Although PC2 accounts for the second largest proportion of variance in the PCA transformation (0.71%), it does not have a significant impact on the performance of the prediction model. PC1 and PC4 exhibit a negative influence on the predictions, whereas PC3 has a contrasting positive effect. Referring to Figure 3, PC1 demonstrates strong spectral loading at wavelengths of 1,210 nm, 1,350 nm, 1,690 nm, 1,930 nm, and 2,229 nm, indicating interactions with hydroxyl (–OH) groups from water and carbohydrates, as well as carbonyl (–C=O) groups that contribute to variations in moisture content [39]. PC4 exhibits a similar absorption pattern, with additional emphasis at 1,403 nm and 1,750 nm, which are associated with hydrogen bonding in the lignocellulose matrix [46]. In contrast, PC3 spans a broader spectral range, with dominant absorption bands at 2,010 nm and 2,100 nm, corresponding to methoxy (–$OCH_3$) groups in lignin and aromatic structure [47]. In the SSF process, white-rot fungi, which are lignolytic and cellulolytic microorganisms, produce enzymes that hydrolyze lignocellulose into simpler compounds. A study by Tuyen et al. [14] demonstrated that white-rot fungi effectively degrade structural carbohydrate components into $CO_2$ and $H_2O$, leading to a reduction in NDF and ADF content. At the molecular level, this degradation involves the cleavage of C–O–C bonds in hemicellulose and lignin, as well as the hydrolysis of β-1,4-glycosidic bonds in cellulose [48].

Additionally, this degradation process generates by-products in the form of organic acids, which alter hydrogen ion concentrations and subsequently increase the pH level [12]. Consequently, NIR spectroscopy can serve as an effective indicator of complex carbohydrate degradation during fermentation, enabling precise and real-time monitoring of fermentation dynamics.

## 4. CONCLUSIONS

This study successfully developed a reliable ML prediction model for intelligent and real-time monitoring of moisture and pH changes during the SSF process of feedstuffs. PCA-transformed NIR spectra facilitate the processing of data and enhance computational efficiency. Overall, the SVR model demonstrated superior performance compared to LightGBM and RF. The SVR model for moisture content achieved an $R^2_p$ of 1.00, $RMSE_p$ of 0.04, $RPD_p$ of 45.00, and $RER_p$ of 217.5, while the pH level model attained an $R^2_p$ of 1.00, $RMSE_p$ of 0.03, $RPD_p$ of 33.33, and $RER_p$ of 116.7. These unusually high $R^2_p$ values are likely influenced by the balanced calibration–prediction set composition, controlled measurement environment, and the use of PCA to minimize noise, and may not directly translate to larger, more heterogeneous datasets. Although SVR yielded superior predictive performance, LightGBM offered notable practical advantages, particularly in terms of computational efficiency. Its leaf-wise tree growth enables rapid training even with high-dimensional spectral data, making it more suitable for real-time or embedded monitoring scenarios where computational resources are limited. Thus, LightGBM provides a balanced trade-off between accuracy and speed in operational settings. The observed spectral features linked to moisture content (around 1,400, 1,900, and 2,100-2,300 nm) and pH variation (1,100-1,500 nm and 2,100-2,300 nm) align with functional group changes (O–H, N–H, C=O) resulting from lignocellulose decomposition by white-rot fungi, explaining the observed increases in both parameters during SSF. This approach establishes a robust foundation for further applications in developing microbial growth monitoring systems within animal feed bioprocesses. Future studies should incorporate completely independent batch validation sets to rigorously assess the robustness and generalizability of the predictive models. Additionally, expanding the model to include further fermentation-related parameters (such as lignin, NDF, and ADF degradation) would provide deeper insight into SSF dynamics and support the development of more comprehensive and efficient process optimization strategies.

## ACKNOWLEDGMENT

This research was supported by the Ministry of Education, Culture, Research, and Technology – Indonesia, through the Doctoral Dissertation Research Grant 2024, as specified in Assignment Agreement Number 094/E5/Pg.02.00.Pl/2024.

## REFERENCES

[1] Manurung, R., Melinda, R., Abduh, M.Y., Widiana, A., Sugorc, I., Suheryadi, D. (2015). Potential use of lemongrass (*Cymbopogon winterianus*) residue as dairy cow feed. Pakistan Journal of Nutrition, 14(12): 919-923. https://doi.org/10.3923/PJN.2015.919.923

[2] Sari, A.F., Manguwardoyo, W., Sugoro, I. (2017). Degradasi ampas dan serai wangi segar (C*ymbopogon nardus* L) dengan metode *in sacco* pada kerbau fistula. In Prosiding Seminar Nasional Teknologi Peternakan dan Veteriner, pp. 118-124. https://doi.org/10.14334/PROS.SEMNAS.TPV-2017-P.119-125

[3] Singh, M., Pandey, N., Dwivedi, P., Kumar, V., Mishra, B.B. (2019). Production of xylose, levulinic acid, and lignin from spent aromatic biomass with a recyclable Brønsted acid synthesized from d-limonene as renewable feedstock from citrus waste. Bioresource Technology, 293: 122105. https://doi.org/10.1016/J.BIORTECH.2019.122105

[4] Vong, S., Yi, T., Net, S., Morm, S., Lunpha, A., Yeanpet, C., Pilajun, R. (2025). The effects of treated dried cassava stem replacement on feed intake, digestibility, rumen fermentation, and blood metabolites of Thai native cattle. Animal Bioscience, 38(3): 501-510. https://doi.org/10.5713/AB.24.0577

[5] Nayan, N., Sonnenberg, A.S.M., Hendriks, W.H., Cone, J.W. (2018). Screening of white-rot fungi for bioprocessing of wheat straw into ruminant feed. Journal of Applied Microbiology, 125(2): 468-479. https://doi.org/10.1111/JAM.13894

[6] Sousa, D., Simões, L., Oliveira, R., Salgado, J.M., Cambra-López, M., Belo, I., Dias, A. (2023). Evaluation of biotechnological processing through solid-state fermentation of oilseed cakes on extracts bioactive potential. Biotechnology Letters, 45(10): 1293-1307. https://doi.org/10.1007/s10529-023-03417-4

[7] Uber, T.M., Backes, E., Saute, V.M.S., da Silva, B.P., Corrêa, R.C.G., Kato, C.G., Seixas, F.A.V., Bracht, A., Peralta, R.M. (2023). Enzymes from basidiomycetes—peculiar and efficient tools for biotechnology. In Biotechnology of Microbial Enzymes, pp. 129-164. https://doi.org/10.1016/B978-0-443-19059-9.00023-2

[8] Harada, A., Sasaki, K., Kaneta, T. (2016). Direct determination of lignin peroxidase released from Phanerochaete chrysosporium by in-capillary enzyme assay using micellar electrokinetic chromatography. Journal of Chromatography A, 1440: 145-149. https://doi.org/10.1016/j.chroma.2016.02.062

[9] Bentil, J.A., Thygesen, A., Mensah, M., Lange, L., Meyer, A.S. (2018). Cellulase production by white-rot basidiomycetous fungi: Solid-state versus submerged cultivation. Applied Microbiology and Biotechnology, 102(14): 5827-5839. https://doi.org/10.1007/s00253-018-9072-8

[10] Lu, X., Zhao, Y., Li, F., Liu, P. (2023). Active polysaccharides from *Lentinula edodes* and *Pleurotus ostreatus* by addition of corn straw and xylosma sawdust through solid-state fermentation. International Journal of Biological Macromolecules, 228: 647-658. https://doi.org/10.1016/j.ijbiomac.2022.12.264

[11] Calleja-Gómez, M., Roig, P., Brnčić, S.R., Barba, F.J., Castagnini, J.M. (2023). Scanning electron microscopy and triple TOF-LC-MS-MS analysis of polyphenols from pef-treated edible mushrooms (*L. edodes*, *A. brunnescens*, and *P. ostreatus*). Antioxidants, 12(12): 2080. https://doi.org/10.3390/antiox12122080

[12] Perwez, M., Al Asheh, S. (2025). Valorization of agro-

industrial waste through solid-state fermentation: Mini review. Biotechnology Reports, 45: e00873. https://doi.org/10.1016/J.BTRE.2024.E00873

[13] Jiang, H., Liu, G., Mei, C., Yu, S., Xiao, X., Ding, Y. (2012). Measurement of process variables in solid-state fermentation of wheat straw using FT-NIR spectroscopy and synergy interval PLS algorithm. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 97: 277-283. https://doi.org/10.1016/J.SAA.2012.06.024

[14] Tuyen, D.V., Phuong, H.N., Cone, J.W., Baars, J.J.P., Sonnenberg, A.S.M., Hendriks, W.H. (2013). Effect of fungal treatments of fibrous agricultural by-products on chemical composition and in vitro rumen fermentation and methane production. Bioresource Technology, 129: 256-263. https://doi.org/10.1016/J.BIORTECH.2012.10.128

[15] He, Q., Chen, H.Z. (2015). Comparative study on occurrence characteristics of matrix water in static and gas double-dynamic solid-state fermentations using low-field NMR and MRI. Analytical and Bioanalytical Chemistry, 407(30): 9115-9123. https://doi.org/10.1007/s00216-015-9077-4

[16] Zimmerleiter, R., Kager, J., Nikzad-Langerodi, R., Berezhinskiy, V., Westad, F., Herwig, C., Brandstetter, M. (2020). Probeless non-invasive near-infrared spectroscopic bioprocess monitoring using microspectrometer technology. Analytical and Bioanalytical Chemistry, 412(9): 2103-2109. https://doi.org/10.1007/s00216-019-02227-w

[17] Jiang, Q., Zhang, M., Mujumdar, A.S., Wang, D. (2023). Non-destructive quality determination of frozen food using NIR spectroscopy-based machine learning and predictive modelling. Journal of Food Engineering, 343: 111374. https://doi.org/10.1016/J.JFOODENG.2022.111374

[18] Fatemi, A., Singh, V., Kamruzzaman, M. (2022). Identification of informative spectral ranges for predicting major chemical constituents in corn using NIR spectroscopy. Food Chemistry, 383: 132442. https://doi.org/10.1016/J.FOODCHEM.2022.132442

[19] Malvandi, A., Kapoor, R., Feng, H., Kamruzzaman, M. (2022). Non-destructive measurement and real-time monitoring of apple hardness during ultrasonic contact drying via portable NIR spectroscopy and machine learning. Infrared Physics Technology, 122: 104077. https://doi.org/10.1016/J.INFRARED.2022.104077

[20] Jiang, Q., Bao, Y., Ma, T., Tsuchikawa, S., Inagaki, T., Wang, H., Jiang, H. (2025). Intelligent monitoring of post-processing characteristics in 3D-printed food products: A focus on fermentation process of starch-gluten mixture using NIR and multivariate analysis. Journal of Food Engineering, 388: 112357. https://doi.org/10.1016/J.JFOODENG.2024.112357

[21] Samadi, Wahyudi, I., Wajizah, S., Zulfahrizal, Munawar, A.A. (2023). Rapid and non-destructive prediction of animal feed nutritive parameters using near infrared spectroscopy and multivariate analysis. International Journal of Design and Nature and Ecodynamics, 18(4): 951-956. https://doi.org/10.18280/ijdne.180422

[22] Samadi, Wahyudi, I., Wajizah, S., Zulfahrizal, Munawar, A.A. (2024). Robust near infrared spectroscopy for rapid and simultaneous determination of fermented cocoa pod husk feed quality attributes. International Journal of Design and Nature and Ecodynamics, 19(2): 379-386. https://doi.org/10.18280/ijdne.190203

[23] Dai, C., Xu, X., Huang, W. Yan, P., Hou, Y., He, R., Ma, H. (2023). Monitoring of critical parameters in thermophilic solid-state fermentation process of soybean meal using NIR spectroscopy and chemometrics. Journal of Food Measurement and Characterization, 17: 576-585. https://doi.org/10.1007/s11694-022-01628-3

[24] Munawar, A.A., Zulfahrizal, Mörlein, D. (2024). Prediction accuracy of near infrared spectroscopy coupled with adaptive machine learning methods for simultaneous determination of chlorogenic acid and caffeine on intact coffee beans. Case Studies in Chemical and Environmental Engineering, 10: 100913. https://doi.org/10.1016/J.CSCEE.2024.100913

[25] Gao, W., Jiang, Q., Guan, Y., Huang, H., Liu, S., Ling, S., Zhou, L. (2024). Transfer learning improves predictions in lignin content of Chinese fir based on Raman spectra. International Journal of Biological Macromolecules, 269: 132147. https://doi.org/10.1016/J.IJBIOMAC.2024.132147

[26] Ma, J., Zheng, B., He, Y. (2022). Applications of a hyperspectral imaging system used to estimate wheat grain protein: A review. Frontiers in Plant Science, 13: 837200. https://doi.org/10.3389/FPLS.2022.837200/PDF

[27] Ao, T.J., Wu, J., Chandra, R., Zhang, H.Y., Yuan, Y.F., Luo, Y.P., Li, D., Liu, C.G., Renneckar, S., Saddler, J. (2025). Influence of hemicellulose and lignin on the effect of drying of cellulose and the subsequent enzymatic hydrolysis. Green Chemistry, 27(29): 8901-8913. https://doi.org/10.1039/d5gc02029h

[28] Lever, J., Krzywinski, M., Altman, N. (2017). Principal component analysis. Nature Methods, 14(7): 641-642. https://doi.org/10.1038/nmeth.4346

[29] Wen, Y., Liu, X., He, F., Shi, Y., Chen, F., Li, W., Song, Y., Li, L., Jiang, H., Zhou, L., Wu, L. (2024). Machine learning prediction of stalk lignin content using fourier transform infrared spectroscopy in large scale maize germplasm. International Journal of Biological Macromolecules, 280: 136140. https://doi.org/10.1016/J.IJBIOMAC.2024.136140

[30] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems, 30: 1-9.

[31] Zlobin, M., Bazylevych, V. (2025). Bayesian optimization for tuning hyperparametrs of machine learning models: A performance analysis in Xgboost. Computer Systems and Information Technologies, 1: 141-146. https://doi.org/10.31891/csit-2025-1-16

[32] Oyedele, O. (2023). Determining the optimal number of folds to use in a k-fold cross-validation: A neural network classification experiment. Research in Mathematics, 10(1): 2201015. https://doi.org/10.1080/27684830.2023.2201015

[33] Sim, J., McGoverin, C., Oey, I., Frew, R., Kebede, B. (2023). Near-infrared reflectance spectroscopy accurately predicted isotope and elemental compositions for origin traceability of coffee. Food Chemistry, 427: 136695. https://doi.org/10.1016/J.FOODCHEM.2023.136695

[34] Qi, X., Wang, S., Fang, C., Jia, J., Lin, L., Yuan, T. (2025). Machine learning and SHAP value interpretation for predicting comorbidity of cardiovascular disease and

cancer with dietary antioxidants. Redox Biology, 79: 103470. https://doi.org/10.1016/J.REDOX.2024.103470

[35] López-Calleja, A.C., Cuadra, T., Barrios-González, J., Fierro, F., Fernández, F.J. (2012). Solid-state and submerged fermentations show different gene expression profiles in cephalosporin C production by *Acremonium chrysogenum*. Journal of Molecular Microbiology and Biotechnology, 22(2): 126-134. https://doi.org/10.1159/000338987

[36] Ebrahimi, I., de Castro, R., Ehsani, R., Brillante, L., Feng, S. (2024). Advancing grape chemical analysis through machine learning and multi-sensor spectroscopy. Journal of Agriculture and Food Research, 16: 101085. https://doi.org/10.1016/J.JAFR.2024.101085

[37] Lanjewar, M.G., Morajkar, P.P., Parab, J. (2025). Robust method for detecting metanil yellow in turmeric: Integrating Vis-NIR spectroscopy and machine learning. Journal of Food Composition and Analysis, 142: 107409. https://doi.org/10.1016/J.JFCA.2025.107409

[38] Dai, C., Xu, X., Huang, W., Yan, P., Hou, Y., He, R., Ma, H. (2023). Monitoring of critical parameters in thermophilic solid-state fermentation process of soybean meal using NIR spectroscopy and chemometrics. Journal of Food Measurement and Characterization, 17(1): 576-585. https://doi.org/10.1007/s11694-022-01628-3

[39] Zheng, R., Jia, Y., Ullagaddi, C., Allen, C., Rausch, K., Singh, V., Schnable, J.C., Kamruzzaman, M. (2024). Optimizing feature selection with gradient boosting machines in PLS regression for predicting moisture and protein in multi-country corn kernels via NIR spectroscopy. Food Chemistry, 456: 140062. https://doi.org/10.1016/J.FOODCHEM.2024.140062

[40] Qiu, G., Lü, E., Lu, H., Xu, S., Zeng, F., Shui, Q. (2018). Single-kernel FT-NIR spectroscopy for detecting supersweet corn (*Zea mays* L. *saccharata sturt*) seed viability with multivariate data analysis. Sensors, 18(4): 1010. https://doi.org/10.3390/S18041010

[41] An, T., Yu, S., Huang, W., Li, G., Tian, X., Fan, S., Dong, C., Zhao, C. (2022). Robustness and accuracy evaluation of moisture prediction model for black tea withering process using hyperspectral imaging. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 269: 120791. https://doi.org/10.1016/J.SAA.2021.120791

[42] Kabir, M.F., Chen, T., Ludwig, S.A. (2023). A performance analysis of dimensionality reduction algorithms in machine learning models for cancer prediction. Healthcare Analytics, 3: 100125. https://doi.org/10.1016/J.HEALTH.2022.100125

[43] Pensupa, N., Jin, M., Kokolski, M., Archer, D.B., Du, C. (2013). A solid-state fungal fermentation-based strategy for the hydrolysis of wheat straw. Bioresource Technology, 149: 261-267. https://doi.org/10.1016/J.BIORTECH.2013.09.061

[44] Vega García, M., Aznarte, J.L. (2020). Shapley additive explanations for $NO_2$ forecasting. Ecological Informatics, 56: 101039. https://doi.org/10.1016/J.ECOINF.2019.101039

[45] Yu, M., Yan, J., Chu, J., Qi, H., Xu, P., Liu, S., Zhou, L., Gao, J. (2025). Accurate prediction of wood moisture content using terahertz time-domain spectroscopy combined with machine learning algorithms. Industrial Crops and Products, 227: 120771. https://doi.org/10.1016/J.INDCROP.2025.120771

[46] Gordobil, O., Herrera, R., Poohphajai, F., Sandak, J., Sandak, A. (2021). Impact of drying process on kraft lignin: lignin-water interaction mechanism study by 2D NIR correlation spectroscopy. Journal of Materials Research and Technology, 12: 159-169. https://doi.org/10.1016/J.JMRT.2021.02.080

[47] Wittner, N., Gergely, S., Slezsák, J., Broos, W., Vlaeminck, S.E., Cornet, I. (2023). Follow-up of solid-state fungal wood pretreatment by a novel near-infrared spectroscopy-based lignin calibration model. Journal of Microbiological Methods, 208: 106725. https://doi.org/10.1016/J.MIMET.2023.106725

[48] Lu, Z., DeJong, S.A., Cassidy, B.M., Belliveau, R.G., Myrick, M.L., Morgan, S.L. (2016). Detection limits for blood on fabrics using attenuated total reflection fourier transform infrared (ATR FT-IR) spectroscopy and derivative processing. Applied Spectroscopy, 71(5): 839-846. https://doi.org/10.1177/0003702816654154