JTEP Jurnal Keteknikan Pertanian,
Vol. 14 No. 1, p 144-161, 2026
P-ISSN 2407-0475 E-ISSN 2338-8439

Available online:
http://journal.ipb.ac.id/index.php/jtep
DOI: 10.19028/jtep.014.1.144-161

Research Article

OPEN ACCESS Check for updates

Rapid Analysis of Fresh Cow Milk Chemical Composition by Using Portable NIR Spectrometer Coupled with Machine Learning

Aminatur Ridho[1], Epi Taufik[1], Cahyo Budiman[1], Diang Sagita[2], Slamet Widodo[3*]

[1]Department of Animal Production and Technology, Faculty of Animal Science, IPB University, Kampus IPB Dramaga, Bogor, Jawa Barat 16680, Indonesia

[2]Research Center for Equipment Manufacturing Technology, National Research and Innovation Agency, South Tangerang, Banten 15314, Indonesia

[3]Division of Biosystem Engineering, Faculty of Engineering and Technology, IPB University, Kampus IPB Dramaga, Bogor, Jawa Barat 16680, Indonesia

*Corresponding author, email: slamet_ae39@apps.ipb.ac.id

Article Info

Submitted:* *21 Juli 2025
Revised:* *27 March 2026
Accepted:* *30 March 2026
Available online:* *22 April 2026
Published:* *March 2026

Keywords:
Chemometrics, machine learning, milk composition, portable NIRS, rapid analysis.

How to cite:
Ridho, A., Taufik, E., Budiman, C., Sagita, D., Widodo, S. (2026). Rapid Analysis of Fresh Cow Milk Chemical Composition by Using Portable NIR Spectrometer Coupled with Machine Learning. Jurnal Keteknikan Pertanian, 14(1): 144-161. https://doi.org/10.19028/jtep.014.1.144-161.

Abstract

Analysis of milk composition is essential for quality assurance and compliance with regulations related to quality standards, yet current tools lack of rapid analysis capability especially for field application. This study investigated the potential use of portable near-infrared spectroscopy (NIRS) combined with chemometrics and machine learning to predict fat, protein and lactose content of fresh cow milk. Spectral data were collected from fresh cow milk samples using a portable device working in the short-wave NIR (740–1070 nm). Samples were obtained from five farms at two locations in the Bogor area during morning and evening milking times. As a reference to develop the predictive model, the fat, protein, and lactose contents were measured using Milkotester Master Eco, which is a standard widely accepted by farmers and the milk industry. Two predictive methods were applied: Partial Least Squares Regression (PLS-R) and machine learning algorithms (i.e. Artificial Neural Network (ANN) and Random Forest (RF)) with various data pre-treatments. The best PLS-R models achieved determination coefficient of prediction (R^2p) values of 0.828 (fat), 0.397 (protein), and 0.384 (lactose). Machine learning models further improved R^2p to 0.901, 0.562, and 0.444, respectively. These findings demonstrate that portable NIRS combined with machine learning enables fast and reliable milk composition analysis, particularly for fat content. However, the prediction performance for protein and lactose is still limited and needs to be further improved.

Doi: https://doi.org/10.19028/jtep.014.1.144-161

1. Introduction

Based on data from the Central Bureau of Statistics or Badan Pusat Statistik (BPS), national milk consumption in Indonesia reached 16.27 kg per capita in 2020 (BPS, 2021). From 2017 to 2021, the

 Available online:
http://journal.ipb.ac.id/index.php/jtep
DOI: 10.19028/jtep.014.1.144-161

average annual milk consumption increased by approximately 4,413,009 tons (Ginting et al., 2023). This trend suggests a continued rise in national milk consumption in the coming years. The growing demand for milk necessitates that producers adopt advanced technology-based equipment. It is essential for producers to evaluate the chemical composition of milk prior to distribution to ensure compliance with the Indonesian national quality standards or Standar Nasional Indonesia (SNI) outlined in SNI 3141.1:2011, which addresses physical, chemical, and microbiological properties (BSNI, 2011).

Traditionally, milk composition analysis has relied on laboratory-based chemical methods, which are often time-consuming and pose environmental risks because of the chemicals involved. Furthermore, traditional methods render samples unusable after the analysis. The use of the Milkotester instrument, which is widely used as a rapid analytical tool for assessing milk composition. However, its bulkiness and reliance on main power as a corded/plug-in device limit its suitability for real-time field applications (Diaz-Olivares et al., 2023).

An alternative method for milk analysis is near-infrared spectroscopy (NIRS), which has been widely applied to analyze the composition of various agricultural products. NIRS operates on the principle of interaction between electromagnetic waves emitted by a spectrometer and the chemical bonds present in organic materials (Kusumiyati et al., 2020). Recent advancements in electronics and manufacturing have facilitated the development of portable NIRS devices that are user-friendly and suitable for field application. Previous studies have evaluated the use of portable NIRS for tasks such as organic milk authentication (Liu et al., 2018) and the prediction of milk quality traits (Guerra et al., 2024).

The evolution of analytical instruments has paralleled advancements in data analysis techniques, particularly with the increasing adoption of machine learning (ML). Machine learning excels at rapidly processing large and complex datasets, often yielding superior results in various tasks, such as classification and prediction. It can also identify patterns that may elude human observers (Situmorang & Yahfizham, 2023). Machine learning has been applied to predict agricultural product quality (Kaya et al., 2020; Sagita et al., 2024; Widyaningrum et al., 2024), food processing (Menichetti et al., 2023), and forecasting food crop yields (Satria et al., 2023). In the dairy sector, Fourier Transform Infrared (FTIR) spectral data combined with machine learning algorithm has been utilized to detect whey cheese contamination in milk (Lima et al., 2022), and NIR spectroscopy has been combined with machine learning techniques to predict blood metabolic profiles in dairy cattle (Giannuzzi et al., 2022).

Given the promising potential of portable devices alongside machine learning capabilities, this study aims to integrate both technologies to develop a predictive system for fresh cow's milk composition. The anticipated outcome is a rapid analysis system that can be employed in field settings

Available online:
http://journal.ipb.ac.id/index.php/jtep
DOI: 10.19028/jtep.014.1.144-161

for quality assurance. The results of this study are expected to be beneficial to various stakeholders, including farmers, dairy processing companies, and consumers.

2. Material and Methods

2.1 Fresh Milk Sample

Fresh milk was obtained from two farms with five locations in the Bogor area. Milk samples were collected at two milking times (morning and evening). Samples of 200 mL were randomly collected from 80 Holstein Frisian cows. The milk samples were placed in plastic bags and stored in a cool box during transportation. The samples were then stored in a freezer (± -18 °C) until data collection. Prior to measurement, samples in plastic bags were thawed at room temperature until they reached ± 25 °C.

2.2 Acquisition of Spectra with Portable NIR Device

The reflectance spectrum was obtained for fresh milk. Milk samples (10 mL) were placed in a beaker. The reflectance of the samples was measured using a SCiO portable NIR spectrometer (Consumer Physics, Inc., USA) equipped with a special sample holder for liquids (Figure 1). Measurements were carried out at room temperature (± 25 °C), and there were no special environmental conditions during the measurements. Milk spectrum measurements were performed three times for each sample at wavelength of 740-1070 nm. In total, 240 spectral datasets were obtained from 80 samples.

Figure 1. Acquisition of spectra data of milk sample using SCiO NIR device.

2.3 Measurement of Reference Data of Fresh Milk Content

As a reference, the content of fresh milk consists of fat, protein, and lactose. This content was measured using a standard instrument, namely the Milkotester milk analyzer Master Eco (Milkotester Ltd, Bulgaria). This instrument has been widely used and accepted, including for academic research purposes, as conducted by Rakib et al. (2023) and Mukhnizan et al. (2024). This instrument has the

Available online:
http://journal.ipb.ac.id/index.php/jtep
DOI: 10.19028/jtep.014.1.144-161

following specifications: fat (range: 0 – 25%, accuracy: ± 0.1%), protein ((range: 2 – 7%, accuracy: ± 0.15%), and lactose (range: 0.01 – 6%, accuracy: ± 0.2%) (Milkotester, 2019). Prior to the measurement, the milk sample (20 mL) was homogenized and poured into a container. The detector was then dipped into the sample and allowed to stand for approximately 60 s. The measurement results will appear on the connected computer screen.

2.4 Data Analysis

Milk content prediction modeling was carried out using two approaches: 1) Partial Least Squares-Regression (PLS-R) with a spectra data including original and pre-treated data, 2) Machine learning algorithms, namely Random Forest (RF) and Artificial Neural Network (ANN) on three types of data (original spectra, latent variable (LV) of the best PLS-R model, and Principal Component (PC) obtained from Principal Component Analysis of the original data spectra). The selection of these methods (i.e., PLS-R, ANN, and RF) was based on their respective characteristics and literature studies that show that these methods are effective for similar cases. PLS-R is the standard method for processing spectral data and is the most widely used method. ANN, a machine learning algorithm, can capture nonlinear relationships in data. RF represents an ensemble learning method that, in many cases, shows excellent performance in predicting chemical content based on spectral data (Ferragina et al. (2015); Soyeurt et al., 2020; Peng et al., 2025).

PLS-R analysis was performed using Unscrambler® X 10.4 software (CAMO, Norway), whereas modeling using RF and ANN was performed using the scikit-learn package in the Python programming language. For machine learning-based modelling, hyperparameter tuning was performed using cross-validation (10-fold) to optimize the model. The hyperparameters used for the RF and ANN models are listed in Table 1. The data were divided into two groups: 56 samples (70%) for training and 24 samples (30%) for validation. Several data pre-treatments were used, including normalization, Savitzky-Golay smoothing, Derivatives-1 and Derivatives-2, Standard Normal Variate (SNV), detrending, and Multiplicative Scatter Correction (MSC). The performance of the developed predictive model was evaluated using several statistical parameters, including the Root Mean Squared Error (RMSE), coefficient of determination (R^2), and ratio of prediction to deviation (RPD), as presented in Equation 1-3.

$$RMSE = \left[\frac{1}{N}\sum_{i=1}^{N}\left(y_{a,i} - y_{p,i}\right)^2\right]^{\frac{1}{2}} \quad (1)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_{a,i}-y_{p,i})^2}{\sum_{i=1}^{N}(y_{a,i}-\bar{y}_i)^2} \quad (2)$$

$$RPD = \frac{SD}{RMSE_p} \quad (3)$$

Available online:
http://journal.ipb.ac.id/index.php/jtep
DOI: 10.19028/jtep.014.1.144-161

where, N is the number of samples, y_a is the actual value, y_p is the predicted value, SD is the standard deviation, $\overline{y_i}$ is mean value. The subscripts c and p in $R^2{}_c$, $R^2{}_p$, $RMSE_c$, and $RMSE_p$ denote the values of the statistical parameters from the calibration and prediction stages, respectively.

Table 1. Hyperparameters used for model development using machine learning algorithms

Algorithm	Hyperparameter	Range value
Random Forest (RF)	N estimators (NE)	25, 50, 75, 100, 125, 150
	Maximum depth (MD)	None, 10, 20, 30
Artificial Neural Network (ANN)	Hidden layer sizes (HSL)	(14), (28), (56), (100), (14,7), (28,7), (56,7), (100,100), (14,14,14), (28,28,28), (56,56,56)
	Maximum iteration (MI)	5000
	Activation function (AF)	identity, tanh, relu
	Solver	lbfgs, sgd
	Alpha (α)	0.0001, 0.001, 0.01
	Learning rate (LR)	Constant

3. Results and Discussion

3.1 Chemical Composition of Fresh Milk Sample

The important components of the dairy products analyzed in this study were protein, fat, and lactose. Based on the results of the analysis of all samples, protein, fat, and lactose data were obtained for each sample based on milking time (morning and evening). The results of this analysis are presented in Table 2. The average fat content in milk differed between milking times, with 2.58 ± 1.74% in the morning and 3.99 ± 2.16% in the evening. Thus, the fat content in the morning was below the Indonesian national standard, which has a minimum fat content of 3%.

Table 2. Results of milk composition with Milkotester

Parameter	Time Milking	Content (%)	SNI 3141.1.2011 (%)
Protein	Morning	3.32±0.17	Minimum 2.8
	Evening	3.28±0.19	Minimum 2.8
Fat	Morning	2.58±1.74	Minimum 3.0
	Evening	3.99±2.16	Minimum 3.0
Lactose	Morning	5.00±0.26	-
	Evening	4.96±0.27	-

Protein content that meets the standard indicates the quality of feed given to livestock. Based on previous research results (Suhendra et al., 2015), the optimal balance of concentrates and forages is 40% concentrate and 60% forage, which improves the milk quality of dairy cows. The fat content in

Available online:
http://journal.ipb.ac.id/index.php/jtep
DOI: 10.19028/jtep.014.1.144-161

milk during afternoon milking was higher by 3.99±2.16 than in morning milking by 2.55±1.74. This is because the milking time interval from morning to evening is shorter (±11 h) than from afternoon to morning milking (±13 h). Milk quality variations in milking intervals cause differences, especially in protein, fat, and dry matter content (Nugraha et al., 2016). Lactose is one of the constituents of dry matter (total solids) in milk, along with other components, such as fat, protein, vitamins, and minerals. According to Maulidina et al. (2021), a decrease in total solid content is followed by a decrease in fat and protein content.

3.2 Spectral Characteristics of Fresh Milk

Fresh milk reflectance data from the sample measurements in the wavelength range of 740-1070 nm are presented in Figure 2. The graph shows peaks and valleys of specific spectra, which indicate the interaction of NIR waves with certain chemical contents in milk samples. NIR waves interact with chemical bonds in organic molecules, such as C-H, N-H, O-H, and S-H (Liu et al., 2018). More specific in the short-wave NIR region, Wu et al. (2008) mentioned that specific wavelengths that correlated and could be used for predicting fat (927 nm, 992 nm, 1018 nm, 1033 nm and 1040 nm) and protein (904 nm, 949 nm, 987 nm, and 1002 nm). A notable decrease in the spectrum at 960-970 nm indicates the presence of O-H or water bonds. The strong influence of water bond absorption may create a bias in predicting the main bonds in the milk composition, which is similar to the results of previous studies (Riu et al., 2020).

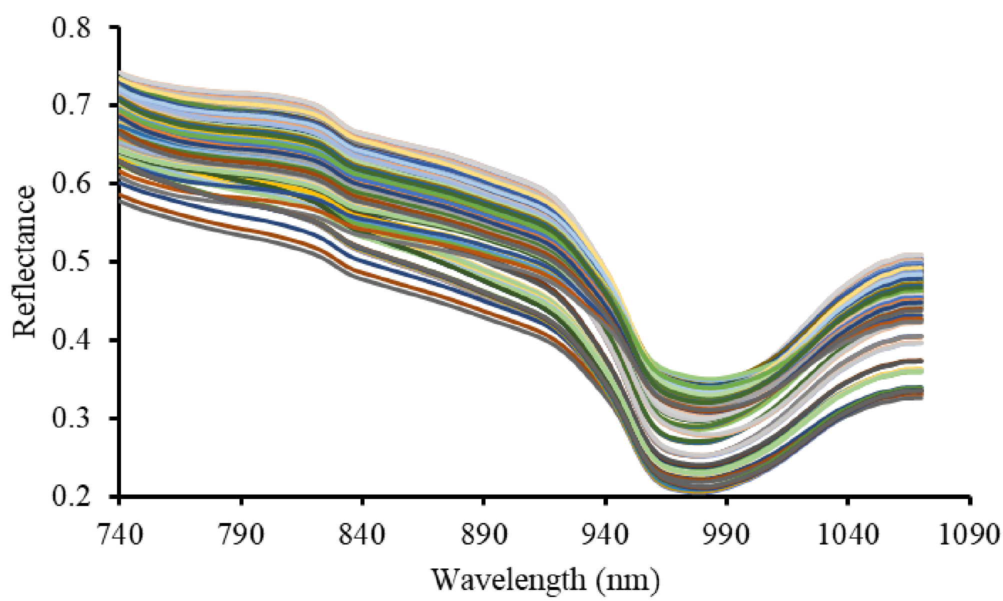

Figure 2. Spectra of sample scanning results with SCiO NIR spectrometer.

Available online:
http://journal.ipb.ac.id/index.php/jtep
DOI: 10.19028/jtep.014.1.144-161

3.3 Prediction of Fat Content

Table 3 shows that SNV pre-treatment provides the best model for predicting fat content using PLS-R at an LV of 8. The value of R^2_c and prediction R^2_p of 0.884 and 0.810 shows that the model has a good performance in predicting fat because it is close to 1 (Ahmad & Sabihah, 2018). The RMSEC and RMSEP values were 0.69 and 0.68, respectively, with a difference of only 0.01. The small difference between RMSEC and RMSEP indicates that the produced model is better and more stable (Lammertyn et al., 2000).

Table 3. Performance of the fat content prediction model using PLS-R.

Pre-treatment	LV	R^2_c	R^2_p	RMSEC	RMSEP	RPD
Raw data	9	0.884	0.828	0.62	0.70	2.50
Savitzky-Golay	9	0.884	0.828	0.62	0.72	2.43
Normalize	8	0.884	0.810	0.62	0.74	2.35
Detrending	6	0.884	0.792	0.64	0.76	2.29
SNV	**8**	**0.865**	**0.828**	**0.69**	**0.68**	**2.57**
MSC	9	0.903	0.449	0.6	0.73	2.40

Table 4. Performance of fat content prediction model using ML

Algorithm	Pre-treatment	Hyperparameter	R^2_c	R^2_p	RMSEC	RMSEP	RPD
RF	Original	MD: 30, NE: 25	0.965	0.709	0.33	0.97	1.87
RF	LV(SNV)	MD: 20, NE: 50	0.971	0.824	0.30	0.76	2.40
RF	PC	MD: 30, NE: 25	0.964	0.814	0.33	0.78	2.34
ANN	Original	AF: relu, α: 0.001, HLS: (56, 7), Solver: lbfgs	0.999	0.769	0.05	0.86	2.09
ANN	**LV(SNV)**	**AF: relu, α: 0.01, HLS: (56, 56, 56), Solver: lbfgs**	**0.999**	**0.901**	**0.01**	**0.57**	**3.19**
ANN	PC	AF: tanh, α: 0.001, HLS: (28, 28, 28), Solver: lbfgs	0.999	0.892	0.01	0.59	3.06

The performance metrics for fat milk prediction using machine learning, as outlined in Table 4, indicate good accuracy and precision. The R^2_c, R^2_p, RMSEC, RMSEP, and RPD values were 0.99, 0.90, 0.01, 0.57, and 3.19, respectively. Additionally, the high RPD value of 3.19 suggests the excellent

Available online:
http://journal.ipb.ac.id/index.php/jtep
DOI: 10.19028/jtep.014.1.144-161

predictive capability and reliability of the model (Williams, 2010). The choice of the Standard Normal Variate (SNV) pre-treatment method for estimating fat content is attributed to the suspicion of high scattering values in the obtained spectrum data. Scattering occurs when the light received by the NIR device does not pass smoothly through the object or surface but is reflected in various directions. The SNV is optimal in this scenario because it effectively mitigates the effects of scattering, enhancing the accuracy of fat content estimation in milk.

3.4 Prediction of Protein Content

Based on the results presented in Table 5, it is evident that the pre-treatment method used for predicting milk protein content, specifically normalization at an LV of 10, yielded suboptimal performance. This conclusion was drawn based on the relatively low values obtained for R2p and RPD, which were 0.397 and 1.25, respectively. In this case, the modest R2p and RPD values suggest that further optimization or alternative pretreatment methods may be necessary to enhance the accuracy and robustness of the protein content prediction model. The obtained results may not be optimal because of the insufficient wavelength range for predicting the protein components in milk. The use of the NIRONE 2.5 MEMS NIR type, with a wavelength range of 2000-2450 nm, provides relatively optimal predictions (Uusitalo et al., 2021).

Table 5. Performance of the protein content prediction model using PLS-R.

Pre-treatment	LV	R^2_c	R^2_p	RMSEC	RMSEP	RPD
Raw data	6	0.221	0.109	0.16	0.17	1.06
Savitzky-Golay	10	0.423	0.053	0.14	0.17	1.04
Normalize	**10**	**0.533**	**0.397**	**0.12**	**0.14**	**1.25**
Detrending	9	0.518	0.152	0.12	0.17	1.05
SNV	9	0.476	0.212	0.13	0.17	1.06
MSC	8	0.449	0.314	0.13	0.16	1.11

The performance of protein milk prediction using machine learning, as presented in Table 6, demonstrates a notable accuracy and precision. The obtained values for R^2_c, R^2_p, RMSEC, RMSEP, and RPD were 0.96, 0.56, 0.03, 0.11, and 1.52, respectively. ANN algorithms, such as those employed in this study, are recognized for their capability to handle complex or imprecise data and effectively tackle unstructured problems (Sunandar & Sutopo, 2024). Other studies have also successfully utilized artificial intelligence/machine learning algorithms, including ANN and RF, for the detection and monitoring of key parameters (i.e., composition, adulteration detection, and indicators of microbial quality) of dairy products (Jiménez-Carvelo et al., 2022; Mhapsekar et al., 2025). However, the significant difference in R^2 values between the calibration (R^2_c) and prediction (R^2_p) of protein content indicates overfitting during the training process. Further analysis is needed to determine the reason for this issue and potential solutions to solve it.

Available online:
http://journal.ipb.ac.id/index.php/jtep
DOI: 10.19028/jtep.014.1.144-161

Table 6. Performance of the protein content prediction model using ML.

Algorithm	Pre-treatment	Hyperparameter	R^2_c	R^2_p	RMSEC	RMSEP	RPD
RF	Original	MD: None, NE: 25	0.913	0.247	0.04	0.15	1.16
RF	LV(Normalize)	MD: 20, NE: 100	0.909	0.332	0.04	0.14	1.23
RF	PC	MD: None, NE: 125	0.921	0.392	0.04	0.13	1.29
ANN	Original	AF: relu, α: 0.001, HLS: (56, 56, 56), Solver: lbfgs	0.982	0.138	0.02	0.16	1.08
ANN	**LV(Normalize)**	**AF: relu, α: 0.01, HLS: (56), Solver: lbfgs**	**0.963**	**0.562**	**0.03**	**0.11**	**1.52**
ANN	PC	AF: relu, α: 0.01, HLS: (28, 28, 28), Solver: lbfgs	0.991	0.504	0.01	0.12	1.02

3.5 Prediction of Lactose Content

Based on Table 7, the best pretreatment for lactose content prediction using PLS-R is detrending at an LV of 9. However, other parameters remained suboptimal, such as R^2c of 0.52, R^2p of 0.38, and RPD of 1.22. Lactose is the most significant component of milk, besides water (Vidyanto et al., 2015). Nevertheless, the prediction accuracy for lactose remains limited, mainly due to significant spectral interference from the strong O–H absorption of water, which overlaps with the characteristic absorption bands of lactose. Unlike fat and protein, which contain distinct functional groups such as ester and amide bonds that produce more specific NIR features, lactose exhibits weaker and less distinguishable spectral signals in the short-wave NIR region than fat and protein. This is similar to the study by Guerra et al. (2024), who did not find the best model for milk lactose prediction.

Table 7. Performance of lactose content prediction model using PLS-R.

Pre-treatment	LV	R^2_c	R^2_p	RMSEC	RMSEP	RPD
Raw data	2	0.040	0.029	0.26	0.24	1.09
Savitzky-Golay	2	0.040	0.029	0.26	0.24	1.09
Normalize	6	0.260	0.058	0.23	0.24	1.08
Detrending	**9**	**0.518**	**0.384**	**0.18**	**0.21**	**1.22**
SNV	8	0.360	0.116	0.21	0.26	1.01
MSC	8	0.360	0.130	0.21	0.26	1.02

Available online:
http://journal.ipb.ac.id/index.php/jtep
DOI: 10.19028/jtep.014.1.144-161

Table 8. Performance of lactose content prediction model using ML.

Algorithm	Pre-treatment	Hyperparameter	R^2_c	R^2_p	RMSEC	RMSEP	RPD
RF	Original	MD: None, NE: 150	0.914	0.272	0.06	0.21	1.18
RF	LV(Detrending)	MD: None, NE: 125	0.903	0.410	0.07	0.19	1.31
RF	PC	MD: 30, NE: 75	0.919	0.369	0.06	0.19	1.27
ANN	Original	AF: alpha, α: 0.001, HLS: (100), Solver: lbfgs	0.995	0.269	0.02	0.21	1.18
ANN	LV(Detrending)	AF: tanh, α: 0.01, HLS: (100, 100), Solver: lbfgs	0.999	0.252	0.01	0.21	1.16
ANN	**PC**	**AF: relu, α: 0.01, HLS: (56), Solver: lbfgs**	**0.998**	**0.444**	**0.01**	**0.18**	**1.35**

The performance metrics for lactose milk prediction using machine learning, as illustrated in Table 8, exhibited better performance than PLS-R. The obtained values for R^2_c, R^2_p, RMSEC, RMSEP, and RPD were 0.99, 0.44, 0.01, 0.18, and 1.35, respectively. As with protein estimation, the significant difference in R^2 values between the calibration (R^2_c) and predictions (R^2_p) indicates overfitting during the training process. The scatter plots of the best model for predicting the fat, protein, and lactose content in milk using ANN and RF machine learning models are presented in Figures 3, 4, and 5, respectively.

JTEP Jurnal Keteknikan Pertanian,
Vol. 14 No. 1, p 144-161, 2026
P-ISSN 2407-0475 E-ISSN 2338-8439

Available online:
http://journal.ipb.ac.id/index.php/jtep
DOI: 10.19028/jtep.014.1.144-161

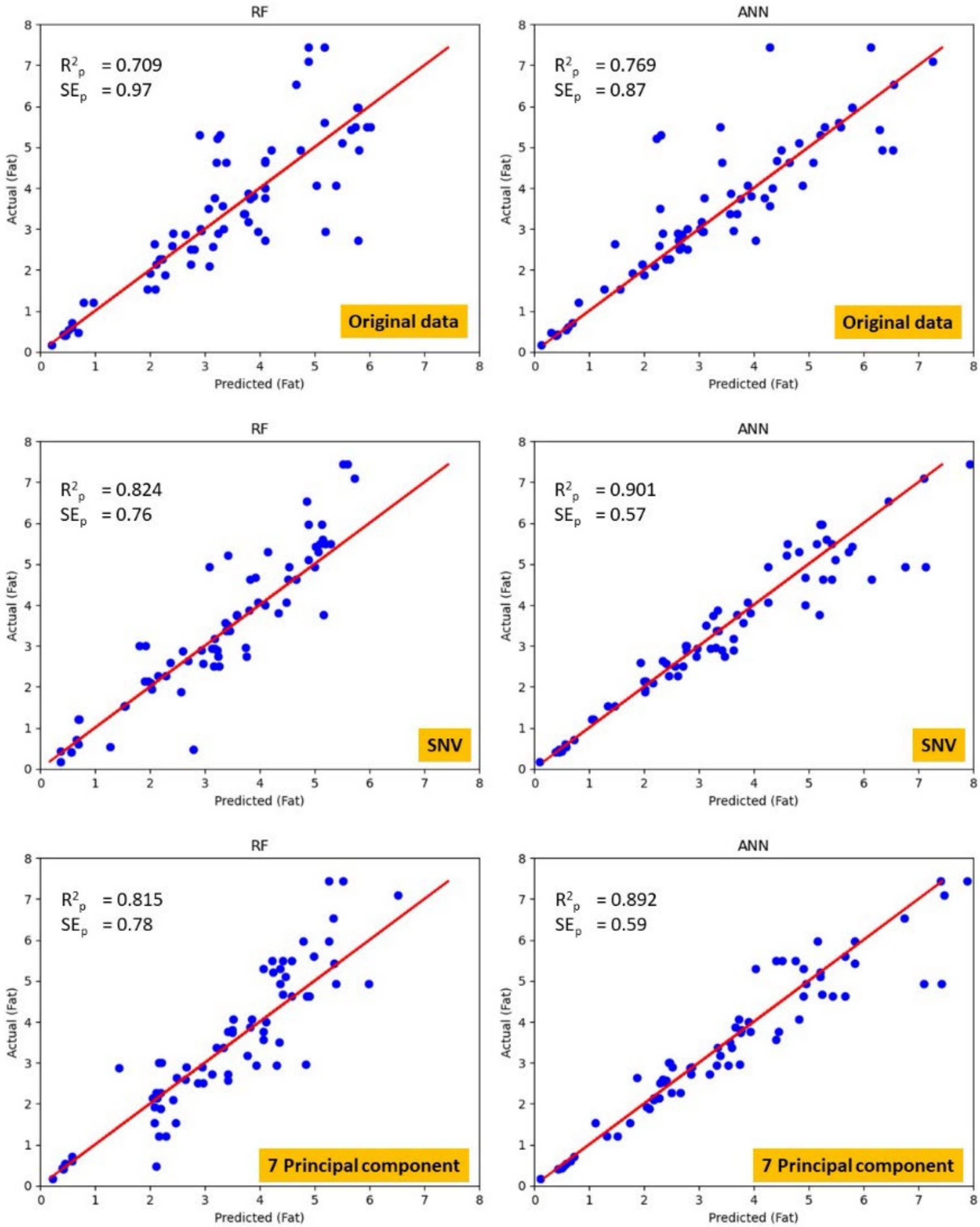

Figure 3. Performance prediction of milk fat using RF and ANN model.

Available online:
http://journal.ipb.ac.id/index.php/jtep
DOI: 10.19028/jtep.014.1.144-161

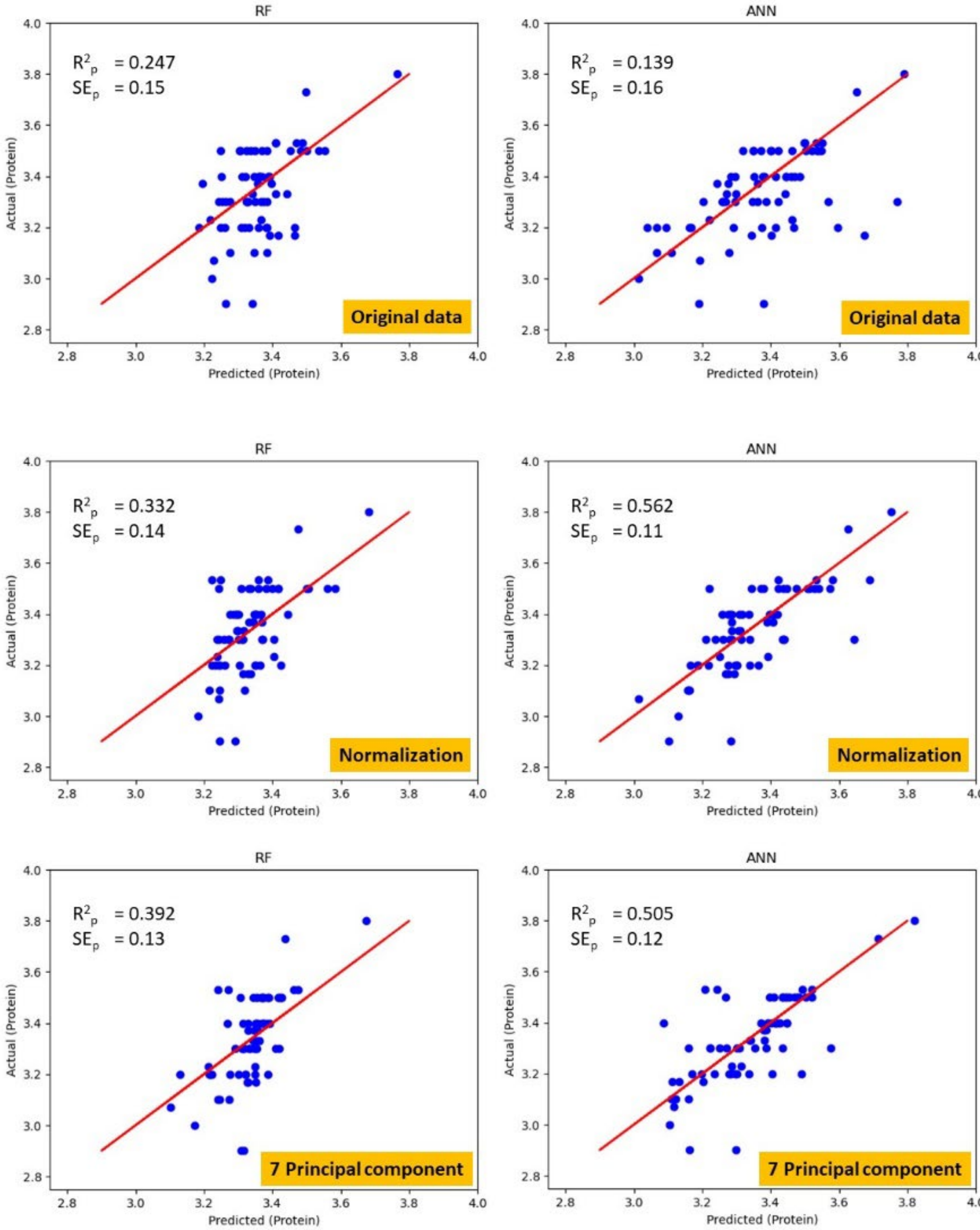

Figure 4. Performance prediction of milk protein using RF and ANN model.

Available online:
http://journal.ipb.ac.id/index.php/jtep
DOI: 10.19028/jtep.014.1.144-161

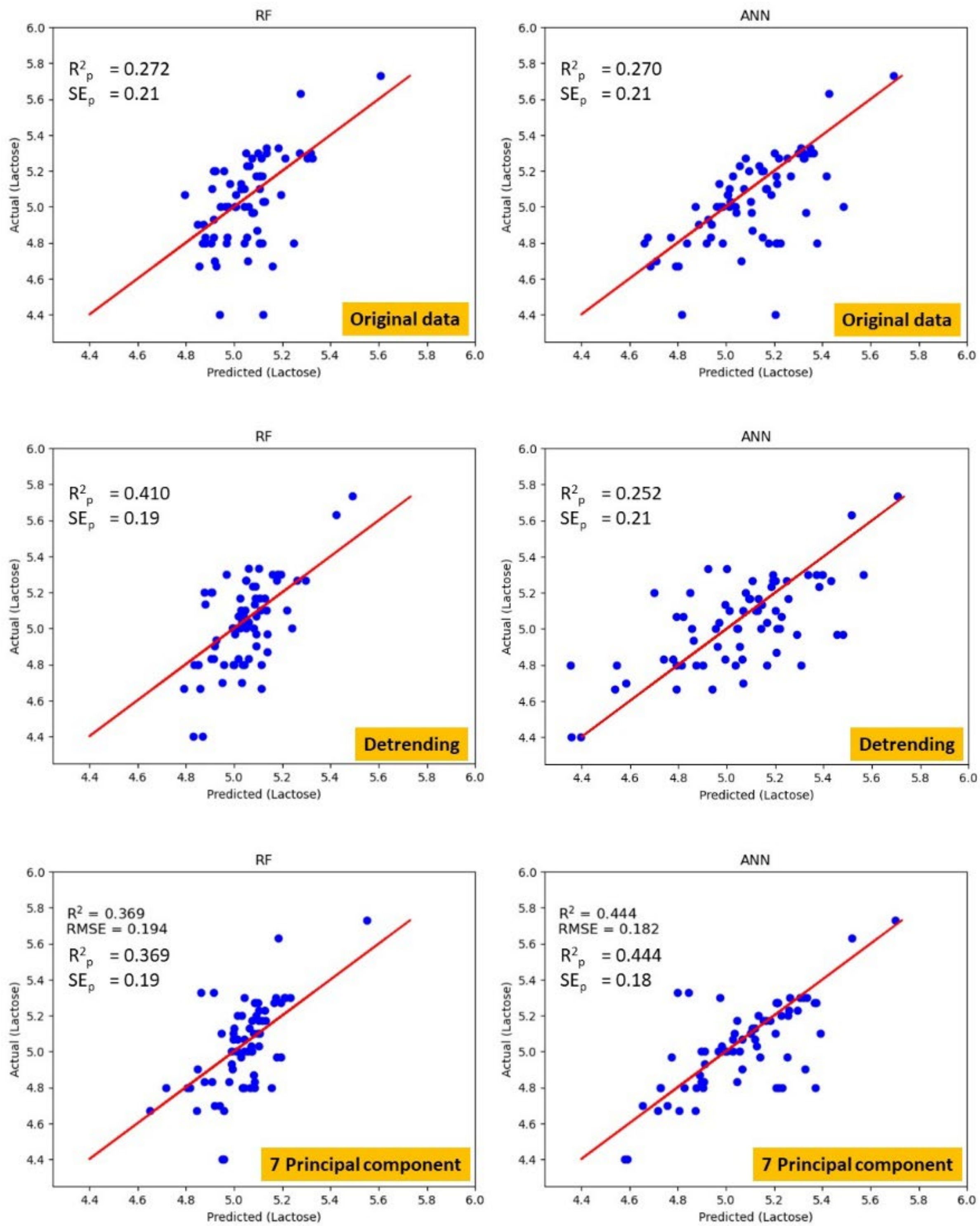

Figure 5. Performance prediction of milk lactose using RF and ANN model.

Overall, the utilization of machine learning presents a marked improvement in the prediction results over the traditional PLS-R method. This underscores the effectiveness of machine learning algorithms in enhancing the predictive capabilities and accuracy of milk composition analyses. However, of the best models obtained for the three components (i.e., protein, fat, and lactose), only the fat content prediction model performed well and was acceptable for all the samples. Referring to the model performance assessment based on the RPD value proposed by Williams (2010), the fat content prediction model with an RPD of 3.0-3.4 was considered a good model that could be used in quality control. Meanwhile, the protein and lactose content prediction models with an RPD value of 0.0-1.9 were considered very poor models and therefore not recommended for use.

During model development, hyperparameter tuning optimization was performed, which appeared to be effective for fat content estimation. However, the indications of overfitting in the estimation require further analysis. Several reasons for this occur, including the relatively limited variation or range of data. The data in Table 1 show that the fat content has a higher range and variation than protein and lactose. Another possible cause is the relatively small amount of data compared with the number of variables used as predictors. According to Ying (2023), several possible ways to address the problem of overfitting include early stopping and reducing the network size during machine learning model training to reduce noise, and data expansion by adding more data, which may be necessary for fine-tuning the machine learning hyperparameters.

In terms of practicality, the use of a smartphone-based portable spectrometer, as used in this study, offers distinct advantages over conventional spectrometers. In terms of hardware, such a system requires only a primary sensor (i.e., a miniaturized spectrometer) and a simple communication device, such as Bluetooth. This allows low-cost system development. Integration with smartphones, which are now widely owned by almost everyone and have high computing capabilities, mobile internet connectivity, and various other built-in sensors, further supports the development of the system. The use of machine learning-based prediction models, which require higher computational power, is no longer a significant problem. The use of machine learning opens up opportunities for developing prediction systems using low-cost, low-resolution sensors. The superiority of machine learning in learning data patterns, which is superior to conventional analysis methods, can compensate for the weaknesses of low-cost, low-resolution sensor-based systems.

4. Conclusion

NIRS technology, operating within a short-wave wavelength range of 740-1070 nm, has emerged as a promising alternative for the classification of the chemical composition of fresh milk samples. The use of machine learning, such as ANN and Random Forest, can enhance the performance of portable NIR SCiO in predicting milk composition, including fat, protein, and lactose content, compared to chemometric methods such as PLS-R. The ANN model achieved the best predictions, with R^2p values of 0.901, 0.562, and 0.444 for the fat, protein, and lactose contents, respectively. The RMSEP values for these predictions were 0.57, 0.11, and 0.18, and the RPD values were 3.19, 1.52, and 1.35, respectively. Despite the limited performance of protein and lactose prediction, combining machine learning methods with portable NIR applications offers a quick and effective way to analyze the fat content of milk. This strategy has the potential to significantly improve the procedures used to evaluate the quality of milk in various applications, including research projects and dairy industry operations.

5. Acknowledgement

The authors would like to thank the Ministry of Education, Culture, Research, and Technology (KEMENDIKBUDRISTEK) of the Republic of Indonesia for funding this research under the Strategic Research Collaboration Program or Program Kolaborasi Penelitian Strategis (KATALIS) with the contract number 34756/IT3.D10/PT.01.03/P/B/2024.

6. AI Writing Statement

No AI or AI-assisted technologies were used in the preparation of this manuscript.

7. References

[BSNI] Badan Standarisasi Nasional Indonesia. 2011. SNI.01-3141-2011: Susu Segar. Jakarta (ID). Badan Standarisasi Nasional.

[BPS] Badan Pusat Statistik. 2021. Jumlah Rata-Rata Konsumsi Susu Tahun 2020. Jakarta (ID). Badan Pusat Statistik.

Ahmad, U., & Sabihah. (2018). Prediction of ripeness parameters of melon fruit using near-infrared spectroscopy. Jurnal Ilmu Pertanian Indonesia, 23(3), 183–189. https://doi.org/10.18343/jipi.23.3.183

Diaz-Olivares, J. A., van Nuenen, A., Gote, M. J., Díaz, V. F., Saeys, W., Adriaens, I., & Aernouts, B. (2023). Near-infrared spectra dataset of milk composition in transmittance mode. Data in Brief, 51. https://doi.org/10.1016/j.dib.2023.109767.

Ferragina, A., de los Campos, G., Vazquez, A. I., Cecchinato, A., & Bittante, G.. (2015). Bayesian regression models outperform partial least squares methods for predicting milk components and technological properties using infrared spectral data. Journal of Dairy Science 98(11), pp. 8133-8151.https://doi.org/10.3168/jds.2014-9143.

Giannuzzi, D., Mota, L. F. M., Pegolo, S., Gallo, L., Schiavon, S., Tagliapietra, F., Katz, G., Fainboym, D., Minuti, A., Trevisi, E., & Cecchinato, A. (2022). In-line near-infrared analysis of milk coupled with machine learning methods for the daily prediction of blood metabolic profile in dairy cattle. Scientific Reports, 12(1). https://doi.org/10.1038/s41598-022-11799-0.

Available online:
http://journal.ipb.ac.id/index.php/jtep
DOI: 10.19028/jtep.014.1.144-161

Ginting, D. V. B., Salsabila, F., Maharani, I., Arzako, M. A., & Sitanggang, N. O. (2023). Analisis pengaruh produksi dan konsumsi terhadap impor susu sapi di indonesia tahun 2017-2021. Investama : Jurnal Ekonomi dan Bisnis (Journal of Economic and Business, 9(1), 57–66. https://doi.org/10.56997/investamajurnalekonomidanbisnis.v9i01.811.

Guerra, A., De Marchi, M., Niero, G., Chiarin, E., & Manuelian, C. L. (2024). Application of a short-wave pocket-sized near-infrared spectrophotometer to predict milk quality traits. Journal of Dairy Science, 107(6). https://doi.org/10.3168/jds.2023-24302.

Jiménez-Carvelo, A. M., Bikrani, S., Nechar, M., Souhail, B., & Cuadros-Rodríguez, L.. (2022). Machine learning–based chemometric methods for quality and authentication of milk and dairy products. in Dairy Foods. Editor(s): da Cruz, A. G., Ranadheera, C. S., Filomena Nazzaro, Mortazavian, A. M. Woodhead Publishing, 261-280, doi: 10.1016/B978-0-12-820478-8.00002-X.

Kaya, A., Keçeli, A. S., Catal, C., & Tekinerdogan, B. (2020). Sensor failure tolerable machine learning-based food quality prediction model. Sensors (Switzerland), 20(11). https://doi.org/10.3390/s20113173.

Kusumiyati, Arip Munawar, A., & Suhandy, D. (2020). Prediction of vitamin c, titratable acidity, and soluble solids content of mango fruits using near-infrared reflectance spectroscopy. Jurnal Teknologi Pertanian, 21(3), 145–154.

Lammertyn, J., Peirs, A., De Baerdemaeker, J., & Nicolaï, B. (2000). Light penetration properties of NIR radiation in fruit with respect to non-destructive quality assessment. Postharvest Biology and Technology, 18(2). https://doi.org/10.1016/S0925-5214(99)00071-X.

Lima, J. S., Ribeiro, D. C. S. Z., Neto, H. A., Campos, S. V. A., Leite, M. O., Fortini, M. E. de R., de Carvalho, B. P. M., Almeida, M. V. O., & Fonseca, L. M. (2022). A machine learning proposal method to detect milk tainted with cheese whey. Journal of Dairy Science, 105(12). https://doi.org/10.3168/jds.2021-21380.

Liu, N., Parra, H. A., Pustjens, A., Hettinga, K., Mongondry, P., & van Ruth, S. M. (2018). Evaluation of portable near-infrared spectroscopy for organic milk authentication. Talanta, 184. https://doi.org/10.1016/j.talanta.2018.02.097.

Maulidina, A., Taufik, E., & Atabany, A. (2021). Performance analysis of fresh milk outbound logistics at north bandung cattle farmer cooperative lembang. Jurnal Ilmu Produksi Dan Teknologi Hasil Peternakan, 9(2), 95–101. https://doi.org/10.29244/jipthp.9.2.95-101.

Menichetti, G., Ravandi, B., Mozaffarian, D., & Barabási, A. L. (2023). Machine learning prediction of the degree of food processing. Nature Communications, 14(1). https://doi.org/10.1038/s41467-023-37457-1.

Mhapsekar, R., Kilbane, D., Davy, S., Abraham, L., Fenelon, M., & O'shea, N. (2025). A systematic review of the internet of things and artificial intelligence applications in milk qualitymonitoring and analysis. International Journal of Dairy Technology, 78(3), doi: 10.1111/1471-0307.70049

Milkotester. (2019). Milk Analyzer Master Eco Brocure. Available at https://milkotester.com/collections/our-products/products/master-eco.

Mukhnizan, M. N., Abdullah, N.A., Hussain, N. M., Ghani, H. S. M., & Ibrahim, N. S. (2024). Malaysian Journal of Applied Sciences 9(1): 64-74. https://doi.org/10.37231/myjas.2024.9.1.389.

Nugraha, B. K., Salman, L. B., & Hernawan, E. (2016). Study of solid fat content, protein, and solid non-fat Holstein dairy cow in the morning and afternoon milking at KPSBU Lembang. Jurnal Unpad, 5.

Peng H, Yi L, Fan X, Zhang J, Gu Y, & Wang S. (2025). Near-infrared spectroscopy assisted by random forest for predicting the physicochemical indicators of yak milk powder. Food Chem. 478:143555. doi: 10.1016/j.foodchem.2025.143555.

Rakib, M. R. H., Messina, V., Gargiulo, J. I., Pathirana, I. N., Lyons, N. A., & Garcia, S. C. (2026). Comparison between in-house competitive and commercial ELISA for the detection of heat shock protein 70 in milk, blood, and saliva of dairy cows. Journal of Dairy Science, in press, https://doi.org/10.3168/jds.2025-27811.

Riu, J., Gorla, G., Chakif, D., Boqué, R., & Giussani, B. (2020). Rapid analysis of milk using low-cost pocket-size NIR spectrometers and multivariate analysis. Foods, 9(8). https://doi.org/10.3390/foods9081090.

Sagita, D., Mardjan, S. S., Suparlan, Purwandoko, P. B., & Widodo, S. (2024). Low-cost IoT-based multichannel spectral acquisition systems for roasted coffee beans evaluation: Case study of roasting degree classification using machine learning. Journal of Food Composition and Analysis, 133(June), 106478. https://doi.org/10.1016/j.jfca.2024.106478.

Satria, A., Badri, R. M., & Safitri, I. (2023). Prediksi hasil panen tanaman pangan sumatera dengan metode machine learning. Digital Transformation Technology, 3(2), 389–398. https://doi.org/10.47709/digitech.v3i2.2852.

Situmorang, S., & Yahfizham. (2023). Analisis kinerja algoritma machine learning dalam deteksi anomali jaringan. Jurnal Matematika Dan Ilmu Pengetahuan Alam, 1(4), 258–269. https://doi.org/10.59581/konstanta.v1i4.1722.

Soyeurt H, Grelet C, McParland S, Calmels M, Coffey M, Tedde A, Delhez P, Dehareng F, & Gengler N. (2020). A comparison of 4 different machine learning algorithms to predict lactoferrin content in bovine milk from mid-infrared spectra. Journal of Dairy Science, 103(12), https://doi.org/10.3168/jds.2020-18870.

Available online:
http://journal.ipb.ac.id/index.php/jtep
DOI: 10.19028/jtep.014.1.144-161

Suhendra, D., T. Anggiati, G., Sarah, S., F. Nasrullah, A., Thimoty, A., & W. C. Utama, D. (2015). Tampilan kualitas susu sapi perah akibat imbangan konsentrat dan hijauan yang berbeda. Jurnal Ilmu-Ilmu Peternakan, 25(1), 42–46. https://doi.org/10.21776/ub.jiip.2015.025.01.06

Sunandar, N., & Sutopo, J. (2024). Utilization of artificial neural network in rice plant disease classification using leaf image. International Journal of Research In Science & Engineering, 42, 1–10. https://doi.org/10.55529/ijrise.42.1.10.

Uusitalo, S., Diaz-Olivares, J., Sumen, J., Hietala, E., Adriaens, I., Saeys, W., Utriainen, M., Frondelius, L., Pastell, M., & Aernouts, B. (2021). Evaluation of mems nir spectrometers for on-farm analysis of raw milk composition. Foods, 10(11). https://doi.org/10.3390/foods10112686

Vidyanto, T., Sudjatmogo, & Sayuthi, S. M. (2015). Tampilan produksi, berat jenis, kandungan laktosa dan air pada susu sapi perah akibat interval pemerahan yang berbeda. Animal Agriculture Journal, 4(2), 200–203.

Widyaningrum, Purwanto, Y. A., Widodo, S., Supijatno, & Iriani, E. S. (2024). Rapid assessment of vanilla (*Vanilla planifolia*) quality parameters using portable near-infrared spectroscopy combined with random forest. Journal of Food Composition and Analysis, 133(March), 106346. https://doi.org/10.1016/j.jfca.2024.106346.

Williams, P. (2010). The RPD Statistic: A Tutorial Note. NIR news 25(1):22-26. doi:10.1255/nirn.1419

Wu, D., Yong He, Y., & Feng, S. (2008). Short-wave near-infrared spectroscopy analysis of major compounds in milk powder and wavelength assignment. Analytica Chimica Acta, 610, 232–242. doi:10.1016/j.aca.2008.01.056.

Ying, X. (2019). An Overview of Overfitting and Its Solutions. J. Phys.: Conf. Ser. 1168 022022. doi: 10.1088/1742-6596/1168/2/022022.