# *Self-learning modules for spectra evaluation*

**Dissertation for obtaining the Doctoral Degree of Natural Science**

**(Dr. rer. nat.)**


**Faculty of Natural Science**

**University of Hohenheim**

Institute of Food Science and Biotechnology

Department of Process Analytics and Cereal Science

Submitted by

*Pegah Sadeghi Vasafi*

*Born on July 14th. 1992 in Esfahan/Iran*


*Stuttgart 2022*

| | |
|---|---|
| Dean: | Prof. Dr. rer. nat. Uwe Beifuss |
| 1. First reviewer: | Prof. Dr. rer. nat. Bernd Hitzmann |
| 2. Second reviewer: | Prof. Dr.-Ing. Jörg Hinrichs |
| Submitted on: | 09.09.2021 |
| Oral examination on: | 11.2.2022 |

This work was accepted by the Faculty of Natural Sciences at the University of Hohenheim on 23.12.2021 as “Dissertation for Obtaining the Doctoral Degree of Natural Sciences”.

## Contents

## Acknowledgements

This thesis was created from August 2019 to December 2021 at the department of Process Analytics and Cereal Science of the University of Hohenheim. I want to show my gratitude to all the people who allowed me to write it.

First and foremost, I want to express my gratitude to Prof. Dr. Bernd Hitzmann, my supervisor, for affording me a pleasant opportunity to complete my PhD at the University of Hohenheim. This PhD would not have been possible without his mentoring and helpful advice.

Next, I'd like to express my gratitude to my family, who have always supported and guided me in my endeavors. This thesis would not have been feasible without them and their long-term support.

In the end, my deep appreciation goes out to my colleagues in the department of Process Analytics and Cereal Science.

## Co-authors

With the knowledge and agreement of the supervisor, Professor Dr. Bernd Hitzmann, this work has been partially published earlier. Part of the research presented in this thesis was done in collaboration with co-authors from the University of Hohenheim.

**Anomaly detection during milk processing by autoencoder neural network based on near-infrared spectroscopy**

By Pegah Sadeghi Vasafi, Olivier Paquet-Durand, Kim Brettschneider, Jörg Hinrichs, Bernd Hitzmann. Published in Food Engineering. Volume 299, page 110510, June 2021 (DOI: 10.1016/j.jfoodeng.2021.110510)

Olivier Paquet-Durand assisted in implementing the autoencoder algorithm.

**Establishing a novel procedure to detect deviations from standard milk processing by using online Raman spectroscopy**

By Pegah Sadeghi Vasafi, Jörg Hinrichs, Bernd Hitzmann. Published in Food Control, volume 131, page 108442. January 2022 (DOI: 10.1016/j.foodcont.2021.108442).

**Comparison of various classification techniques for supervision of milk processing**

By Pegah Sadeghi Vasafi, Bernd Hitzmann. Published in Engineering in Life Sciences. October 2021, in press (DOI: 10.1002/elsc.202100098).

## Publication lists

## Peer-reviewed publications

- Pegah Sadeghi Vasafi, Olivier Paquet-Durand, Kim Brettschneider, Jörg Hinrichs, Bernd Hitzmann
  Anomaly detection during milk processing by autoencoder neural network based on near-infrared spectroscopy. Food Engineering 299 (2021): 110510.

- Pegah Sadeghi Vasafi, Jörg Hinrichs, Bernd Hitzmann
  Establishing a novel procedure to detect deviations from standard milk processing by using online Raman spectroscopy. Food Control 131 (2021): 108442.

- Pegah Sadeghi Vasafi, Bernd Hitzmann
  Comparison of various classification techniques for supervision of milk processing. Engineering in Life Sciences (2021) in press.

## Oral Presentation

- **DECHEMA - 21.09.2021**

Pegah Sadeghi Vasafi, Bernd Hitzmann

Autoencoder neural network for supervision of milk processing

# 1. Preface

## 1.1. Summary

Monitoring milk processing is an essential task as it affects the quality and safety of the final product. The aim of this investigation was to develop and analyse the self-learning system for the supervision of the processing of milk. In the self-learning evaluation module, several algorithms for data analysis of near infrared (NIR) and Raman spectra was implemented for the prediction of sample quality and safety.

In the first part of this thesis, the use of NIR spectroscopy for controlling milk processing was investigated. For this reason, a high-quality quartz flow cell with a 1 mm pathlength including temperature controlling option for liquids was implemented. For sample preparation, UHT-milk with 1.5 % fat content was measured at 5 °C and considered as the reference milk. Samples with various changes such as added water and cleaning solution, different fat content and temperature as well as milks from various suppliers were investigated as the modified samples. A data set from reference and modified samples was obtained with NIR measurements. In this study, first Savitzky-Golay derivative with second polynomial order and window size of 15 was applied. It was compared with the usefulness of raw spectrum and also the combination of raw and first derivative spectrum. For the self-learning sector, an autoencoder neural network was employed. Within this thesis, it was shown that the trained autoencoder using first derivative spectra was capable to detect 5 % added water and 9 % cleaning solution in the milk. However, by using the combination spectra, the difference of 0.1 % in fat concentration was perfectly recognized. These two procedures were able to detect milks from different suppliers and difference of 10 °C in the measurement temperature.

Another part of this work was done using Raman spectroscopy. The aim of this part was to check if the previous result can be improved. In this step, the circulation method was again employed the same as in the previous part. However, because of the heat introduced to the sample by the laser using in Raman spectroscopy and the length of plastic tubes which can be affected by the temperature of the laboratory, the measurement temperature was kept at 10 °C. 1.5 % fat UHT-

milk was utilized as the reference sample. Milks with various changes such as different fat contents, various measurement temperatures and added water or cleaning solution were investigated as the modified samples. In this investigation, not only the autoencoder but also some chemometric models were utilized with the purpose of anomaly detection. Principal component analysis (PCA) was investigated to check if the various samples can be categorized separately. In addition, two chemometric modelling techniques such as principal component regression and Gaussian process regression were tested to check the ability for change detection. By using the results obtained by PCA, a sufficient categorization of various samples was not achieved. While the PCR did not present a promising prediction as the related $R^2$ was 0.7, Gaussian process regression with $R^2$ of 0.97 predicted the changes almost perfectly. The trained autoencoder and Gaussian process regression both were able to define 5 % water and cleaning solution, difference of 0.1 % fat content, and variation of 5 °C in the measurement temperature. In comparison between the autoencoder and Gaussian process regression, it should be mentioned that the Gaussian process regression was capable to determine more abnormal signals than the autoencoder, however, it must be trained with all the possible changes. In contrast, the autoencoder can be trained once just with reference signals and used in online monitoring properly.

As the final part and to detect which type of anomalies happened during the milk processing, several classification approaches such as linear discriminant analysis, decision tree, support vector machine, and k nearest neighbour were utilized. While decision trees and linear discriminant analysis failed to effectively characterize the various types of anomalies, the k nearest neighbor and support vector machine presented promising results. The support vector machine presented an accuracy of 81.4 % for test set, while the k nearest neighbor showed an accuracy of 84.8 %. As a result, it is reasonable to assume that both algorithms are capable of classifying various groups of data accurately. It can help the milk business figure out what's going wrong during the processing of milk.

In general, Raman spectroscopy produced better findings than NIR spectroscopy, because the typical absorption bands of milk components in NIR spectrometers may be impacted by high water absorption combined with substantial light scattering by fat globules. Additionally, the

autoencoder as self-learning system was capable of correctly detecting changes during milk processing, however, classification algorithms can aid in obtaining more details.

## 1.2. Zusammenfassung

Die Überwachung der Milchverarbeitung ist eine wesentliche Aufgabe, da sie die Qualität und Sicherheit des Endprodukts beeinflusst. Das Ziel dieser Untersuchung war die Entwicklung und Analyse eines selbstlernenden Systems zur Überwachung der Milchverarbeitung. In dem selbstlernenden Auswertungsmodul wurden verschiedene Algorithmen zur Datenanalyse implementiert, um die Qualität und Sicherheit der Proben mit Hilfe spektroskopischer Methoden vorherzusagen.

Im ersten Teil dieser Arbeit wurde der Einsatz der Nahinfrarot-Spektroskopie (NIR) zur Kontrolle der Milchverarbeitung untersucht. Zu diesem Zweck wurde eine hochwertige Quarzdurchflusszelle mit einer Schichtdicke von 1 mm und einer Temperiermöglichkeit für Flüssigkeiten eingesetzt. Zur Probenvorbereitung wurde UHT-Milch mit 1,5 % Fettgehalt bei 5 °C gemessen und als Referenzmilch betrachtet. Als modifizierte Proben wurden Proben mit verschiedenen Veränderungen wie Wasser- und Reinigungsmittelzusatz, unterschiedlichem Fettgehalt und Temperatur sowie Milch von verschiedenen Lieferanten untersucht. Mit NIR-Messungen wurde ein Datensatz von Referenz- und modifizierten Proben gewonnen. In dieser Studie wurde die erste Savitzky-Golay-Ableitung mit zweiter Polynomordnung und einer Fenstergröße von 15 verwendet. Sie wurde mit der Auswertegüte des Rohspektrums und auch der Kombination aus Roh- und erstem Ableitungsspektrum verglichen. Für den selbstlernenden Bereich wurde ein neuronales Netz als Autoencoder eingesetzt. Im Rahmen dieser Arbeit wurde gezeigt, dass der trainierte Autoencoder unter Verwendung der ersten Ableitung in der Lage war, 5 % zugesetztes Wasser und 9 % Reinigungslösung in der Milch zu erkennen. Durch die Verwendung der Kombinationsspektren wurde auch der Unterschied von 0,1 % in der Fettkonzentration perfekt erkannt. Diese beiden Verfahren waren in der Lage, Milch von verschiedenen Lieferanten und einem Unterschied von 10 °C bei der Messtemperatur zu erkennen.

Ein weiterer Teil dieser Arbeit wurde mit der Raman-Spektroskopie durchgeführt. Ziel dieses Teils war es, zu prüfen, ob das vorherige Ergebnis verbessert werden kann. In diesem Schritt wurde wieder die gleiche Zirkulationsmethode wie im vorherigen Teil verwendet. Wegen der Wärme, die durch den Laser bei der Raman-Spektroskopie in die Probe eingebracht wird, und der Länge der Kunststoffrohre, die durch die Temperatur im Labor beeinflusst werden kann, wurde die Messtemperatur jedoch bei 10 °C gehalten. Als Referenzprobe wurde UHT-Milch mit 1,5 % Fett verwendet. Milch mit verschiedenen Veränderungen wie unterschiedlichen Fettgehalten, verschiedenen Messtemperaturen und Zusatz von Wasser oder Reinigungslösung wurde als modifizierte Probe untersucht. In dieser Untersuchung wurden nicht nur der Autoencoder, sondern auch einige chemometrische Modelle zur Erkennung von Anomalien eingesetzt. Die Hauptkomponentenanalyse (PCA) wurde untersucht, um zu prüfen, ob die verschiedenen Proben separat kategorisiert werden können. Darüber hinaus wurden zwei chemometrische Modellierungstechniken wie die Hauptkomponentenregression und die Gaußsche Prozessregression getestet, um die Fähigkeit zur Erkennung von Veränderungen zu prüfen. Mit den Ergebnissen der PCA konnte keine ausreichende Kategorisierung der verschiedenen Proben erreicht werden. Während die Hauptkomponentenregression (PCR) keine vielversprechende Vorhersage lieferte, da das zugehörige $R^2$ bei 0,7 lag, sagte die Gaußsche Prozessregression mit einem $R^2$ von 0,97 die Veränderungen nahezu perfekt voraus. Sowohl der trainierte Autoencoder als auch die Gaußsche Prozessregression waren in der Lage, 5 % Wasser und Reinigungslösung, einen Unterschied von 0,1 % Fettgehalt und eine Variation der Messtemperatur von 5 °C zu detektieren. Im Vergleich von Autoencoder und der Gauß'schen Prozessregression ist zu erwähnen, dass die Gauß'sche Prozessregression in der Lage war, mehr anormale Signale zu bestimmen als der Autoencoder, allerdings muss sie mit allen möglichen Änderungen trainiert werden. Im Gegensatz dazu muss der Autoencoder nur einmal mit Referenzsignalen trainiert und kann dann für die Online-Überwachung verwendet werden. Als letzter Teil und um zu erkennen, welche Art von Anomalien während der Milchverarbeitung auftraten, wurden verschiedene Klassifizierungsansätze wie lineare Diskriminanzanalyse, Entscheidungsbaum, Support Vector Machine und K Nearest Neighbour verwendet. Während die Entscheidungsbäume und die lineare Diskriminanzanalyse nicht in der Lage waren, die verschiedenen Arten von Anomalien effektiv zu

charakterisieren, lieferten die K Nearest Neighbour und die Support Vector Machine Methode vielversprechende Ergebnisse. Die Support Vector Machine wies eine Genauigkeit von 81,4 % für den Testsatz auf, während die K Nearest Neighbour Methode eine Genauigkeit von 84,8 % ergab. Daher kann man davon ausgehen, dass beide Algorithmen in der Lage sind, verschiedene Datengruppen genau zu klassifizieren. Dies kann der Milchwirtschaft helfen, herauszufinden, was bei der Verarbeitung von Milch falsch läuft.

Im Allgemeinen lieferte die Raman-Spektroskopie bessere Ergebnisse als die NIR-Spektroskopie, da die typischen Absorptionsbanden der Milchbestandteile in NIR-Spektrometern durch eine hohe Wasserabsorption in Kombination mit einer erheblichen Lichtstreuung durch Fettkügelchen beeinträchtigt werden können. Darüber hinaus war der Autoencoder als selbstlernendes System in der Lage, Veränderungen während der Milchverarbeitung korrekt zu erkennen, jedoch können Klassifizierungsalgorithmen helfen, mehr Details zu erhalten.

# 2. Introduction and Outline

## 2.1 Introduction

### 2.1.1. Milk

Milk, as a complex natural food matrix, includes all of the nutrients required for life to exist. It contains 90 % water, varying levels of fat, protein, and carbohydrates, as well as trace minerals (Varnam & Sutherland 1994). The composition of milk can directly affect the quality and safety of products and therefore, the contents of fat, protein and carbohydrate in milk must be labelled on commercial products (Qin et al. 2017). The appearance, flavor, and aroma of milk are all directly related to its quality from the consumer's perspective. Milk's hue is caused by fat globules scattering reflected light, and its density is related to the number and size of particles. Although appearance is not a measurable criterion, it is critical for quality control; nonetheless, there is no federal standard for it. It is suggested to select white, clean, and no debris as standard appearance. Regarding to the standard regulations, the temperature of milk must never exceed 7 °C and the best option is storing the samples at 5 °C or less. The fat content in milk is varied based on types of milk including the whole milk, reduced fat milk, and skimmed milk. Recent advancements in dairy processing have resulted in increased product safety and quality. Ultra-pasteurization processes and aseptic packaging technologies, in particular, have provided the industrial user with products that have a longer shelf life (Chandan, 2011). The public's attention has recently been drawn to the quality and safety of products due to regular allegations of milk safety issues. As a result, strategies for evaluating and controlling the process are needed in the dairy business. Monitoring processing levels would be critical because the various processes of milk production have a substantial impact on the quality and safety of final products. Consequently, the demands for techniques which can evaluate or control milk quality or safety have been dramatically increased (He et al. 2019). For processing the milk, most refrigerated products are ultra-pasteurized by heating to 125 - 137.8 °C for 2 - 5 seconds and packaged in sterilized cartons in clean atmosphere. For ambient storage, UHT milk is treated at 135 - 148.9 °C for 4 - 15 seconds, followed by aseptic packaging. Homogenization which can reduce the size of

particles is implemented at the temperatures higher than 37 °C. It pumps the milk at high pressure through a small orifice in order to reduce the size of particles. For low fat products or products in which high viscosity is desired, single stage homogenization is employed which means the pressure drop is engineered over a single valve. On the other hand, dual stage homogenization is implemented for fluids with high fat contents or whenever low viscosity is needed. In the first stage the product is subjected to high pressure of 13.8 MPa which results in breakdown of the particle size diameter to an average of less than 1 μm. Then the product goes through the second stage of 3.5 MPa to break the clusters of globules formed in the first stage. Finally, the products are cooled rapidly to 4.4 °C, packed in the proper packages, and stored in the proper places (Chandan, 2011). In this research, various procedures were tested in order to develop a method that can monitor the production of milk and detect the changes which can happen during the process as fast as possible.

### 2.1.2. Change detection

Monitoring various steps of milk production is an essential task as it corresponds to quality and safety of final products. These days, the final product is analyzed to be approved for the quality and safety. Therefore, if unusual changes happen during the production, they can be detected after completion of process which cause wasting time and money. To solve such an effective problem, online monitoring of the process is highly suggested which can helps companies to detect changes immediately and avoid suffering. Thereby, as the aim of this research, developing a detection method based on spectroscopy combined with evaluation techniques was investigated.

### 2.1.3. Near infrared spectroscopy (NIR)

NIR technology is one of the advanced non-destructive systems for quality evaluation which is available for a wide range of applications. NIR spectroscopy at different wavelengths ranges (extending from the visible spectrum to the near infrared) has been implemented to determine various quality parameters. In electromagnetic radiation, three wavelength ranges exist: ultraviolet (UV) radiation from 100 to 380 nm, visible (VIS) radiation from 380 to 780 nm, and

infrared (IR) radiation above 780 nm. The near infrared (NIR) portion of the IR covers wavelengths ranging from 780 nm to 2500 nm. Broad band absorptions related primarily to overtones and combinations of vibrational modes involving C–H, O–H, N–H, and S–H chemical bonds are present in the spectra obtained using this approach (Osborne, 2000). Reflectance spectra give information about both the color and brightness of an object which are widely used as criteria for evaluating product quality (Krivoshiev et al., 2000). Many studies have confirmed the ability of NIR spectroscopy in the food science to predict the main chemical components such as water, protein and fat. It was reported by several researches that NIR spectroscopy could be used as a proper tool for detecting additives in milk powder, liquid milk, and in infant formula (Balabin and Smirnov, 2011). Kawasaki et al. (2008) suggested that the NIR spectroscopic system can be used to assess milk quality in real-time in an automatic milking system. They added that such a system could provide dairy farmers with information on milk quality and physiological or health condition of an individual cow which can give a feedback control for optimizing dairy farm management. By using the system, dairy farmers will be able to produce high-quality milk and precision dairy farming will be realized. In this contribution, NIR was used to simulate the on-line processing of milk. For this reason, a special flow cell including temperature controlling option for liquids was implemented. Therefore, the flow cell was connected by plastic tubes to the source of milk in a cold-water bath and the temperature was kept stable during the process. As a result of this construction, NIR measured the milk during the circulation in various processing steps. The purpose behind it was to create a new technique based on NIR spectroscopy to determine anomalies during the milk processing.

### 2.1.4. Raman spectroscopy

Raman spectroscopy which is based on an inelastic scattering effect was first documented by Raman and Krishnan (Raman and Krishnan, 1928). When a sample is irradiated by a high-intensity monochromatic light, a few of the scattered photons exchange energy with the sample molecules, so the direction and frequency of the scattered photons also change. Such inelastic scattering is called Raman scattering. The majority of scattered photons, on the other hand, do not exchange energy with the molecules in the sample; as a result, these scattered photons

change only the direction of propagation with the same frequency as the excitation photons, and this type of elastic scattering is known as Rayleigh scattering. When the energy of photons changes, they will shift to shorter or larger wavelength, and Stokes or anti-Stokes line can be appeared in spectra (Lord, 1990). Stokes scattering has a substantially higher intensity than anti-Stokes scattering, and it is used in the application of Raman spectroscopy in food analysis (Morris 2006). When the polarizability of a molecule changes during vibration, Raman signals can be seen (Dijkstra et al. 2005). The frequency shifts of dispersed light are revealed by spectral bands in the Raman spectrum. As a result, different spectral bands represent various chemical bonds and functional groups in samples. Raman spectroscopy can be utilized not only for qualitative and structural study of a sample, but also for quantitative determination, because the band intensity is linearly proportional to the concentration of the examined molecule (Yang & Ying 2011). Raman spectroscopy which have been moderately employed for milk analysis has been developed into various analysis technologies such as surface-enhanced Raman spectroscopy (SERS), Fourier-transform (FT) Raman spectroscopy, micro-Raman spectroscopy, near infrared (NIR) Raman spectroscopy, and offset Raman spectroscopy (SORS) (He et al. 2019). Because the Raman signal is inversely proportional to the excitation wavelength, shorter laser excitation wavelengths can produce stronger Raman signals. Fluorescence, on the other hand, can easily interfere with it at shorter excitation wavelengths. Milk analysis has been done using Raman spectroscopy. Moros et al. (2007) implemented FT-Raman spectroscopy to quantify fat content in milk powder. Stefanov et al. (2010) investigated the effects of spectra acquisition temperature conditions for predicting odd-chain and branched-chain fatty acids in milk utilizing a FT-Raman spectroscopy. Also, Hou et al. (2016) employed SERS method for quantitative analysis of milk protein. FT-Raman spectroscopy in combination with partial least discriminant analysis was implemented by Rodrigues et al. (2016) for evaluating and classifying various milk powder samples according to the lactose state and the addition of maltodextrin. In this contribution, to simulate the on-line processing procedure, the quartz flow cell was connected to the source of milk by plastic tubes. Thereby, Raman spectrometer measured the milk during the circulation and the related spectra for further investigations were acquired. It was used with the aim of anomaly detection in the milk processing steps. Several reference and abnormal spectra were

measured by Raman in order to establish a novel method which can distinguish anomalies during the production of milk.

### 2.1.5. Pre-processing

As there are huge baseline shifts and noises in spectra with broad wavelength ranges while analyzing them, choosing the right pre-treatment procedure is crucial (Cen et al., 2006). Pre-processing is employed to obtain effective results in milk analysis and eliminate noise or any irrelevant information**.** For instance, smoothing or Savitzky-Golay filtering techniques are commonly used to improve signal-to-noise ratio (Zhang & Henson, 2007), while baseline correction algorithms can be employed for removing fluorescence background in Raman spectra. Moreover, normalization algorithms (standard normal variate, multiplicative scatter correction, peak normalization) are applied for evaluating samples. More approaches and their functionality are detailed in several literatures (Lohumi et al., 2017; De Luca et al.,2015; Schulze et al., 2005).

### 2.1.6. Principal component analysis and regression

Principal component analysis (PCA) is the most popular multivariate statistical technique used by most of the scientific disciplines. PCA can extract the dominant patterns in data in terms of a complementary set of score and loading plots. The purpose behind it is to extract the necessary information, reduce the dimensionality of data set by keeping the most important information which has decreasing variance with increasing components, and represent it as a set of new orthogonal variables called principal components. Finally, it analyzes the structure of the observations and the variables and shows the pattern of similarity. The first PC presents the largest possible variance of observations while the second one is calculated with respect to the constraint of being orthogonal to the first component. The values of these new variables for the observations are called factor scores which can be interpreted geometrically as the projections of the observations regarding to the principal components (Abdi et al., 2010). Principal component regression (PCR) is a regression analysis technique which is based on PCA. In details, PCR is used for estimating the unknown regression coefficients in a standard linear regression model. In PCR, instead of regressing the dependent variable on the explanatory variables directly,

the principal components of the explanatory variables are used as regressors (Næs et al., 1988). In this work, PCA was implemented for classification of samples including reference and abnormal samples. In addition, PCR was employed in order to predict the difference of standard and abnormal samples. In this case, the reference samples were put equal to zero and abnormal ones equal to one to check if a modified sample (which has some deviations from standard one) can be detected from standards.

### 2.1.7. Autoencoder neural network as self-learning tool

Supervised and unsupervised learning algorithms are generally two types of machine learning techniques, which are normally utilized in sentiment analysis. In supervised learning technique, the dataset is labeled and trained to obtain a reasonable output (Gautam and Yadav, 2014). While unsupervised learning process do not need any labeled data, hence they cannot be processed at ease (Tripathy et al.,2016). Self-learning tools can be considered as tools that can be trained with example samples and consequently they are ready to detect abnormal and normal samples. An autoencoder neural network can be seen as a good example of self-learning tools categorised in the unsupervised learning techniques which can apply back propagation algorithms. The back-propagation neural network is a universal approximator given sufficient hidden units, multilayer feedforward sigmoidal network architectures can approximate virtually any function of interest to any desired degree of accuracy (Sharma et al., 2007). The autoencoder is also a feed forward neural network presenting that information moves in only one direction, from input neurons through the hidden neurons and finally reach the output neurons, thus there is no loop or cycle in it. It tries to set the output values equal to inputs by compressing the inputs to the hidden layers called encoding and decompressing again to the outputs called decoding. While the tangent sigmoid transfer function was employed for hidden neurons, the identity transfer function was implemented for the output neurons. The tangent sigmoid function is given in the equation (1):

$$f(x) = \frac{2}{1+e^{-2x}} - 1 \qquad (1)$$

While x denotes weighted sum of the inputs (Vasafi et al., 2021).

Figure 1 presents the structure of autoencoder simply.

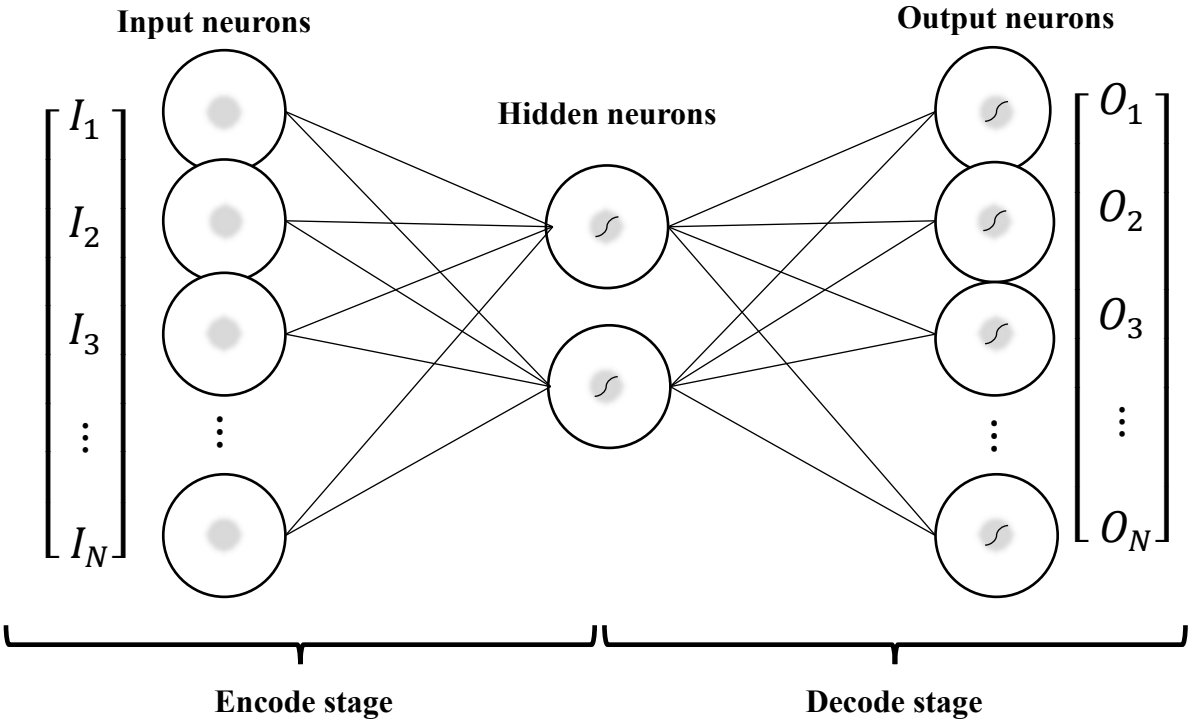


**Figure 1**. The autoencoder structure. N is equal to the number of wavenumber channels.

For anomaly detection during the milk processing, the autoencoder can be trained by spectra from reference samples, thus it can learn to reproduce the reference spectra. While the autoencoder faces the spectra with some deviations which is different from reference sample, it is not able to reproduce the spectra. Therefore, by creating a higher autoencoder error, it can imply the anomalies in the sample. MSE (mean squared error) is used as the autoencoder error which shows average of difference between output and input spectrum for each individual wavelength. Mean squared error is mentioned as equation (2):

$$\mathrm{MSE} = (\vec{x}_{out} - \vec{x}_{in})(\vec{x}_{out} - \vec{x}_{in})^T/n \qquad (2)$$

$\vec{x}_{in}$ and $\vec{x}_{out}$ are the spectra as vectors for the input and output of the autoencoder respectively and n is represented the number of wavenumber channels.

Next, the upper boundary of confidential was calculated in order to define the limit of detection which means that each sample shows the higher autoencoder error than the limit of detection is

considered as the abnormal sample includes a deviation from standard. Although, confidence interval was computed for training set by the formula mentioned below:

Confidence Interval = $\overline{MSE} \pm t(\alpha, f) \times s$ (3)

$\overline{MSE}$ is the average of autoencoder errors (MSE) of all training spectra, $t$ is t distribution, $f$ degree of freedom, and α significance level. In addition, s is equal to standard deviation of autoencoder errors. A significance level of 5 % was utilized (Vasafi et al., 2021).

Generally, by doing so, the autoencoder would be able to detect deviations from standard milk processing which helps companies to be aware of production problems as fast as possible.

### 2.1.8. Gaussian process regression

Gaussian processes have attracted many attentions for statistical data analysis as it shows good predictive performance and analytical properties. Normally, multivariate calibration models have been developed using regression-based techniques. A Bayesian non-parametric regression technique, namely the Gaussian process regression is more flexible and practical than parametric models such as feed-forward neural networks (Chen at al., 2007; Chaurasia et al., 2019). Gaussian process is a distribution over the space of functions where any subset has a Gaussian distribution, therefore it is mainly specified by its mean and covariance function. This algorithm is not useful for applications with large datasets since it does not scale with the number of data points (Ranganathan et al., 2010). In the framework of Gaussian process, wide variety of covariance functions can be employed, subject to the requirement that a valid covariance function must always result in a positive definite covariance matrix for the targets. In a Bayesian model, the covariance function depends mainly on different hyperparameters, which are themselves given prior distributions. These hyperparameters can control the amount of noise in a regression model, the degree to which various input variables are relevant, and the magnitudes of various additive components of a model. The posterior distribution of these hyperparameters will be concentrated on values that are proper for the data that was actually observed. In contrast to the elaborate forms for the covariance function described here, the mean function for the Gaussian process (GP) will usually be set to zero. However, it does not mean that this function is

expected to be centered around zero. It implies that mean function does not make an impact in the prediction output and the prediction is only based on the covariance function. Different covariance functions can be constructed by adding and multiplying other covariance functions, since the element-by-element sum or product of any two symmetric, positive semidefinite matrices is also symmetric and positive semidefinite (Bernardo et al., 1998). Three different covariance functions (kernel) which play important role in building a regression model are squared exponential, Matern, and rational quadratic. The widely used squared exponential covariance function is differentiable infinitely, meaning that the GP with this covariance function has mean square derivatives of all orders, and is therefore very smooth (Aye et al., 2017). In this study, the model was created based on spectra obtained from both reference samples and abnormal ones (while reference samples were put equal to zero and abnormal samples were put equal to one). After training phase, a threshold which can split data into normal and abnormal groups was selected. Finally, the test set was checked to detect anomalies in milk samples. If a sample shows higher value than the threshold, it is considered as the abnormal sample. Therefore, by doing so, the anomalies can be distinguished from standard samples.

## 2.2. Outline

Monitoring the steps of milk processing is an essential sector as can assure the safety and quality of final product. Conventional methods using for this reason are time-consuming, laborious, and require complicated preparation levels (He et al., 2019). Thereby, a time-resaved technique to monitor steps based on online measurement data is highly demanded. By doing so, early changes can be detected immediately which can help companies to react promptly and avoid to be suffered. Accordingly, the focus of this study was to develop novel statistical methods for monitoring milk processing using spectroscopic techniques such as near infrared and Raman spectroscopy.

In the first contribution ‘‘**Anomaly detection during milk processing by autoencoder neural network based on near-infrared spectroscopy’’** the main goal was to develop a self-learning tool which can monitor the milk production steps and detect deviations (such as changes in fat or temperature, added water or cleaning solution) from standard final product. As a new method

of unsupervised machine learning, the autoencoder showed a powerful ability for this purpose, thus the use of the autoencoder combined with near infrared (NIR) spectrometer was investigated. To recognize deviations, raw spectra acquired by NIR were compared with their first derivative and combination of both in order to find the best detection method. The autoencoder was trained by 1.5 % fat UHT-milk as a standard sample which was measured at 5 °C and utilized to define every possible change. The results of using first derivative spectra presented that the trained autoencoder was capable to detect 5 % added water and 9 % cleaning solution in the milk. Although, the trained autoencoder with combination spectra, was able to detect a difference of 0.1 % in fat concentration. In addition, various production methods (milks from various suppliers) and difference of 10 °C in the temperature were determined by both procedures. With the proposed method, it was possible to monitor the various steps of milk processing and detect some deviations from standard final product. To improve the detection, in the second publication **''Establishing a novel procedure to detect deviations from standard milk processing by using inline Raman spectroscopy''** use of Raman spectrometer was checked. Data set obtained from Raman was evaluated with various techniques such as principal component analysis and regression, Gaussian process regression, and the autoencoder in order to determine the most suitable procedure of the detection. In this investigation, principal component regression was not capable to predict the deviations properly. Therefore, the Gaussian process regression was utilized as the next option and illustrated the promising results in the detection of 5 % water and cleaning solution as well as 0.1 % difference in fat content and variation of 5 °C in temperature. The similar results were obtained by using the autoencoder neural network. For a comparison between the autoencoder and Gaussian process regression, it can be mentioned that both procedures are worthy, however the autoencoder can be trained once by only standard samples and used immediately for the on-line supervision. The Gaussian process regression has to be trained by standard samples as well as samples with possible changes and it can be mentioned as the disadvantage of this method. All the previous publications were used to define changes during the process. However, in the third publication ''**Comparison of various classification techniques for supervision of milk processing''** the purpose behind it was to determine the type of anomalies happened in the milk processing. Therefore, various

classification methods such as linear discriminant analysis, decision tree, support vector machine, and k nearest neighbour were implemented. Support vector machine and k nearest neighbour presented promising results. They classified reference milks from modified milks correctly and also categorized different groups of anomalies properly.

# 3. Publications

## 3.1. Anomaly detection during milk processing by autoencoder neural network based on near-infrared spectroscopy

By Pegah Sadeghi Vasafi, Olivier Paquet-Durand, Kim Brettschneider, Jörg Hinrichs, Bernd Hitzmann. Published in Food Engineering. Volume 299, page 110510, June 2021.

Journal of Food Engineering 299 (2021) 110510

Contents lists available at ScienceDirect

Journal of Food Engineering

journal homepage: http://www.elsevier.com/locate/jfoodeng

# Anomaly detection during milk processing by autoencoder neural network based on near-infrared spectroscopy

Pegah Sadeghi Vasafi *, Olivier Paquet-Durand, Kim Brettschneider, Jörg Hinrichs, Bernd Hitzmann

ARTICLE INFO

*Keywords:*
Anomaly detection
Autoencoder
Near-infrared spectroscopy

ABSTRACT

Anomaly detection during milk processing (such as changes in fat or temperature, added water or cleaning solution) can assure a satisfactory final product quality, including compositional and hygienic characteristics, as well as adulteration with water. The use of near-infrared (NIR) spectroscopy for change detection in complex dairy matrix is discussed. The autoencoder neural network plays fundamental role in anomaly detection. To evaluate this capability, the raw spectra obtained from NIR as well as first derivative and combination of both were analysed. An autoencoder was trained by 1.5% fat UHT-milk (measured at 5 °C) and applied to detect possible changes happening during the milk processing. The trained autoencoder using first derivative spectra was capable to detect 5% added water and 9% cleaning solution in the milk. Also, with the combination spectra, it was able to recognize a difference of 0.1% in fat concentration. In addition, both procedures were able to detect different production methods (specific procedure of suppliers such as homogenization level or pressure) and difference of 10 °C in the temperature. It can be concluded, that using an autoencoder neural network in combination with near-infrared spectroscopy is a reliable method to monitor the milk processing. By doing so, abnormal changes can be detected early, controlling the process becomes easier and the quality and safety of the product is guaranteed.

## 1. Introduction

There are strict regulations on the monitoring and control of temperature, composition such as fat concentration, hygiene, taste and smell in the milk processing industry in order to assure a satisfactory final product quality (Bylund, 1995). The main control point here is the temperature, refrigerated milk storage tanks must maintain a temperature of 5 °C. From a consumer standpoint, the quality factors associated with milk are appearance, color, aroma, flavor, and mouth feel which are highly dependent on composition. Added water is an adulteration in milk processing and testing the freezing point of milk using a cryoscopy indicates if abnormal amounts of water exist in the load (Chandan, 2011). The regulations on the hygienic production of milk are strong on intention and require milk to be produced to define quality standards based on traditional methodology. As of now, common testing procedures appear to be less frequent and are often more likely to be reactive. Overall, control or even recognition of a potential problem in milk can be difficult (Hillerton, and Berry, 2004). Having reliable and accurate measurements of the states of the process (such as separation, pasteurization and standardization) is one of the essential principles of process control but it cannot be adequately obtained by the single point information provided by conventional sensing techniques and devices (Williams and Beck, 1995; Scott and McCann, 2005; Sharifi and Young, 2011). Reliable and robust sensors are not available for important quality and process variables, such as sensory assessment, micro flora or spoilage; furthermore, available sensors are typically used only in isolated applications and frequently provide insufficient reliability (Hitzmann et al., 2015). Zettel et al. (2016) point out, that especially optical sensors are able to measure process variables without physically touching the products and without the use of any additional reagents. They predict a more widespread application of optical sensors in all branches of food production in the near future.

Near-infrared (NIR) spectroscopy measurements have great potential for evaluating safety and quality of final products rapidly and nondestructively. Rapid measurement with minimal sample preparation is the most important advantage of spectroscopic methods. Food and beverage processing companies are already using optical sensing technologies for quality and process control (Workman et al., 2003; Huang et al., 2008).

* Corresponding author. Department of Process Analytics and Cereal Science (150i), University of Hohenheim, Garbenstr. 23, 70599, Stuttgart, Germany.
*E-mail addresses:* Pegah.sadeghivasafi@uni-hohenheim.de (P.S. Vasafi), O.Paquet-Durand@uni-hohenheim.de (O. Paquet-Durand), kim.brettschneider@uni-hohenheim.de (K. Brettschneider), J.Hinrichs@uni-hohenheim.de (J. Hinrichs), Bernd.Hitzmann@uni-hohenheim.de (B. Hitzmann).

https://doi.org/10.1016/j.jfoodeng.2021.110510
Received 30 September 2020; Received in revised form 19 January 2021; Accepted 21 January 2021
Available online 25 January 2021

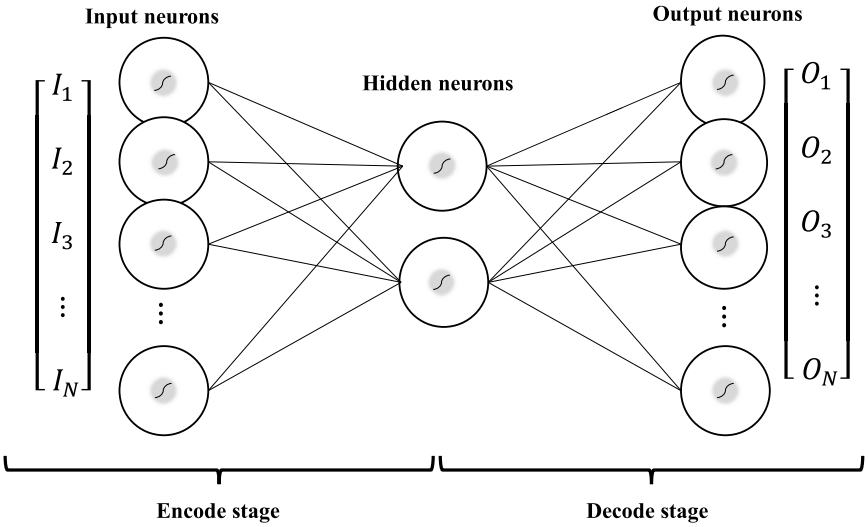


**Fig. 1.** The autoencoder structure. N is equal to the number of wavenumber channels.

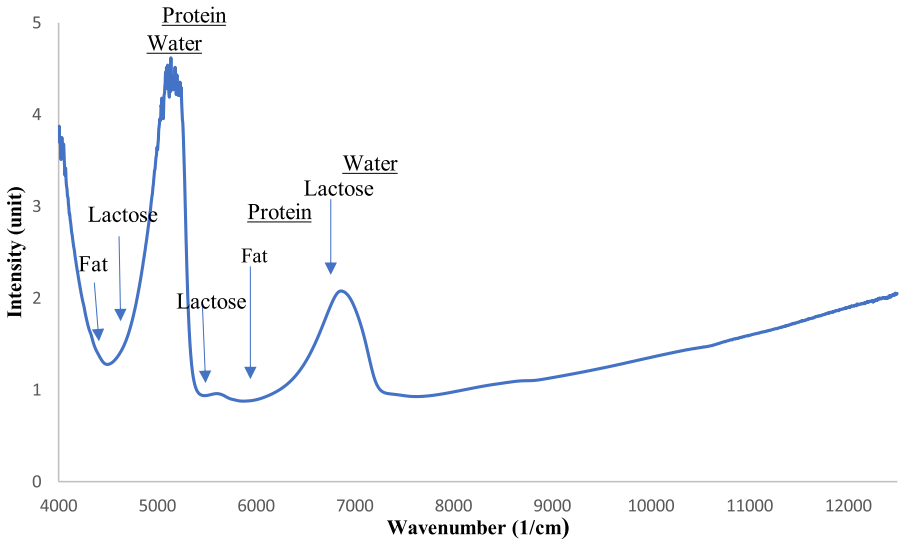


**Fig. 2.** NIR spectra of cow milk acquired from Bruker MPA; the spectral features of some components are mainly distributed around some wavenumbers (underlined) whereas others would appear at the specific wavenumber (arrow).

NIR hyperspectral imaging was investigated for quantitative evaluation of melamine particles in nonfat and whole milk powders (Huang et al., 2016). Using chemometric methods helps to better interpret the results of NIR. It was reported by several research groups that NIR spectroscopy combined with chemometrics could be used as a rapid and accurate method for detecting additives in milk powder, liquid milk, and in infant formula (Mauer et al., 2009; Lu et al., 2009; Balabin and Smirnov, 2011; Fu et al., 2014). Another serious experimental work on the determination of sensory attributes of cheese has been done by using NIR spectrometer based on artificial neural network (ANN). Principle component analysis (PCA) was applied to the dataset (NIR spectra obtained from cheese samples) and was used as input to the ANN. Through ANNs, a reliable prediction was obtained that related the NIR spectrum of a complete set of cheese samples with a complete image of the sensory attributes that described taste, texture, aspect, smell and other relevant sensations. As a result, a very useful tool for quality control with a very low computation cost was obtained (Curto et al., 2020). In other study, NIR spectroscopy, data pretreatment techniques and multivariate data analysis were used to predict fine particle size fraction, dispersibility and bulk density of various milk powder samples, which are believed to have a significant impact on milk powder quality. Predictive models using partial least squares (PLS) regression were developed using NIR spectra and milk powder physical and functional properties, and it was concluded that the PLS models predicted milk powder quality with an accuracy of 88–90% (Khan et al., 2020). Núñez-Sánchez et al. (2016) also employed NIR to evaluate the quality of dairy goat milk and reported that transflectance analysis gave better or similar cross-validation results than reflectance mode. Kawasaki et al. (2008) investigated the NIR sensing system for on-line bovine milk quality assessment in a milking robot. Calibration models for determining three major milk constituents (fat, protein and lactose), somatic cell count and milk urea nitrogen of unhomogenized milk were developed, and the precision and accuracy of the models were validated. The coefficient of determination and standard error of prediction of the validation set for fat were very good. The results of the study suggested that the NIR spectroscopic system can be used to assess milk quality in real-time in an automatic milking system. They added that such a system could provide dairy farmers with information on milk quality and physiological or

**Table 1**
Percentage of autoencoder errors higher than limit of 6.8E-04, and lower than it as well as mean and standard deviation for samples in the prediction set by using raw data without any pre-processing.

| Sample | Percentage of values lower than limit (%) | Percentage of values higher than limit (%) | Average of MSE (unit) | Standard deviation of MSE (unit) |
|---|---|---|---|---|
| 5% cleaning solution | 100 | 0 | 4.4E-04 | 7.0E-05 |
| 9% cleaning solution | 80 | 20 | 5.2E-04 | 1.7E-04 |
| 10% cleaning solution | 70 | 30 | 6.2E-04 | 2.9E-04 |
| 5% water | 86 | 14 | 5.4E-04 | 1.3E-04 |
| 10% water | 0 | 100 | 2.2E-03 | 2.4E-04 |
| 10 °C | 100 | 0 | 4.4E-04 | 6.6E-05 |
| 15 °C | 0 | 100 | 1.6E-03 | 1.0E-04 |
| 1.6% fat | 0 | 100 | 4.7E-03 | 1.2E-03 |
| ESL-milk | 0 | 100 | 7.2E-02 | 5.2E-03 |
| ESL-milk | 0 | 100 | 8.6E-02 | 4.8E-03 |
| UHT-milk | 0 | 100 | 7.3E-03 | 1.2E-03 |
| UHT-milk | 0 | 100 | 1.5E-01 | 3.1E-03 |

health condition of an individual cow which can give a feedback control for optimizing dairy farm management. By using the system, dairy farmers will be able to produce high-quality milk and precision dairy farming will be realized.

As every possible change such as change in fat, temperature, added water or cleaning solution in processing can affect the quality and safety of milk, anomaly detection is an essential task with critical applications in this area. The unsupervised anomaly detection is learning a normal profile (spectra from standard sample) then identifying the samples not conforming to the normal profile as anomalies. Deep autoencoder has been extensively used for anomaly detection. It is a powerful tool to model the high-dimensional data in the unsupervised setting. After training on the normal data, the autoencoder is expected to produce higher reconstruction error for the abnormal inputs than the normal ones, which is adopted as a criterion for identifying anomalies (Gong et al., 2019). The autoencoder is trained by repeated presentation of representative exemplar input-output vector pairs (pair patterns) where the input and output vectors are the same and the network learns through repeated experiences. The back-propagation learning procedure, which can be used in autoencoder training, is based on adjusting the synaptic strengths or weights of the neural connections by a gradient descent search technique in order to minimize a cost function. Here, the squared difference between the network output and input after a given iteration cycle can be used. Weights are usually randomly initialized, and are gradually adjusted towards the minimum of the cost function in the weight space (Eerikäinen et al., 1993).

Autoencoders have not been applied in food science yet, but ANN have already been utilized in some areas. A neural network model based on back-propagation learning has been found useful for prediction of improved dairy yield (305-day milk yield, fat and protein) (Sharma et al., 2007). ANNs have been employed successfully in another study for dairy yield prediction and cow culling classification (Lacroix et al., 1997). Milk production estimates have been successfully obtained in a study by using feedforward ANN (Sanzogni and Kerr, 2001). ANNs have been used for detecting influential variables in the prediction of incidence of clinical mastitis in dairy animals (Heald et al., 2000). A three-layer back-propagation connectionist model has been used for pattern recognition to develop Monterey jack cheese (Sharma et al., 2007), which allows study of real-time control process of cheese production. Also, ANN has been used in modeling of pH and acidity for cheese production (Paquet et al., 2000). ANNs have been successfully used to predict temperature, moisture and fat in slab-shaped foods with edible coatings during deep-fat frying (Mittal and Zhang, 2000). A self-organized network inspired by immune algorithm (SONIA), which improves generalization capability of the back-propagation method, has been reported in a recent study for time-temperature-based food quality prediction system using real meat delivery data. The results have been compared with standard back-propagation and back-propagation with Bayesian regularization algorithm (Widyanto et al., 2005).

Anomaly detection by autoencoder neural networks has been used in some areas like medical science, but it has not been applied in the processing food yet. It was concluded that using autoencoder neural networks for controlling food production processes is a new application of the method which can be useful in determination of changes in the process. By applying autoencoder neural networks, most early changes can be detected and potentially be solved as soon as possible before affecting the final quality. Therefore, the objective of this study was to evaluate anomaly detection (such as changes in fat, temperature, added water or cleaning solution) during the milk processing by using spectroscopy data as well as autoencoder neural network.

## 2. Material and method

### 2.1. Material

1.5% and 3.5% fat milk were purchased from the brand "Weihenstephan", Germany. The milk with the lower fat concentration is called reference milk, because it was used for training the autoencoder. All samples were stored in refrigerators at 5 °C before opening the packages. Concentrated substance "Anti Germ clean A-N 30" was prepared from dairy pilot plant of university of Hohenheim.

### 2.2. Near-infrared spectroscopy

The NIR spectra were acquired with a MPA Multi-Purpose Fourier Transform NIR spectrometer (Bruker, Germany), which was equipped with the software OPUS (Version 7, Bruker, Germany). The spectral range between 12,500 and 4000 $cm^{-1}$ (800–2500 nm) was scanned with a resolution of 2 $cm^{-1}$. Sixty-four (64) absorbance scans were averaged for each sample spectrum ($R_1$). A background spectrum (64 scans) was measured ($R_0$) before starting the measurement which were collected as reference to avoid spectra variation due to unknown factors. A high-quality quartz flow cell of 1 mm pathlength including temperature controlling option for liquids was implemented. Less than 1 mL of milk was needed for each NIR measurement. The measurement of each sample took almost 1 min. For each experiment, the instrument was turned on 30 min before measurement to achieve a stable state. The averaged and background-corrected spectrum was used for further steps. Based on the obtained intensity $R_1$ and $R_0$, absorbance A was calculated by $A = lg(\frac{R_0}{R_1})$. A cold water bath with temperature of 5 °C was implemented to keep the temperature of samples cold during the measurement. A beaker that held cold milk at 5 °C was left in the water bath. The Flow cell was connected by plastic tubes to the source of milk in the beaker. Milk was circulated through the flow cell during the measurement in order to keep the temperature stable and simulate on-line process measurement conditions in a by-pass.

### 2.3. Sample preparation and data collection

There are several factors which may cause some changes in the spectrum of the reference milk (training set). It was investigated how an autoencoder neural network can detect these changes. In order to define different changes such as temperature or fat variation, added water or cleaning solution to the milk and various production methods of milk, a prediction set was set up and feed to the autoencoder to investigate its performance. Raw and derivative spectra as well as the combination of both were used to investigate which spectral pre-processing can improve

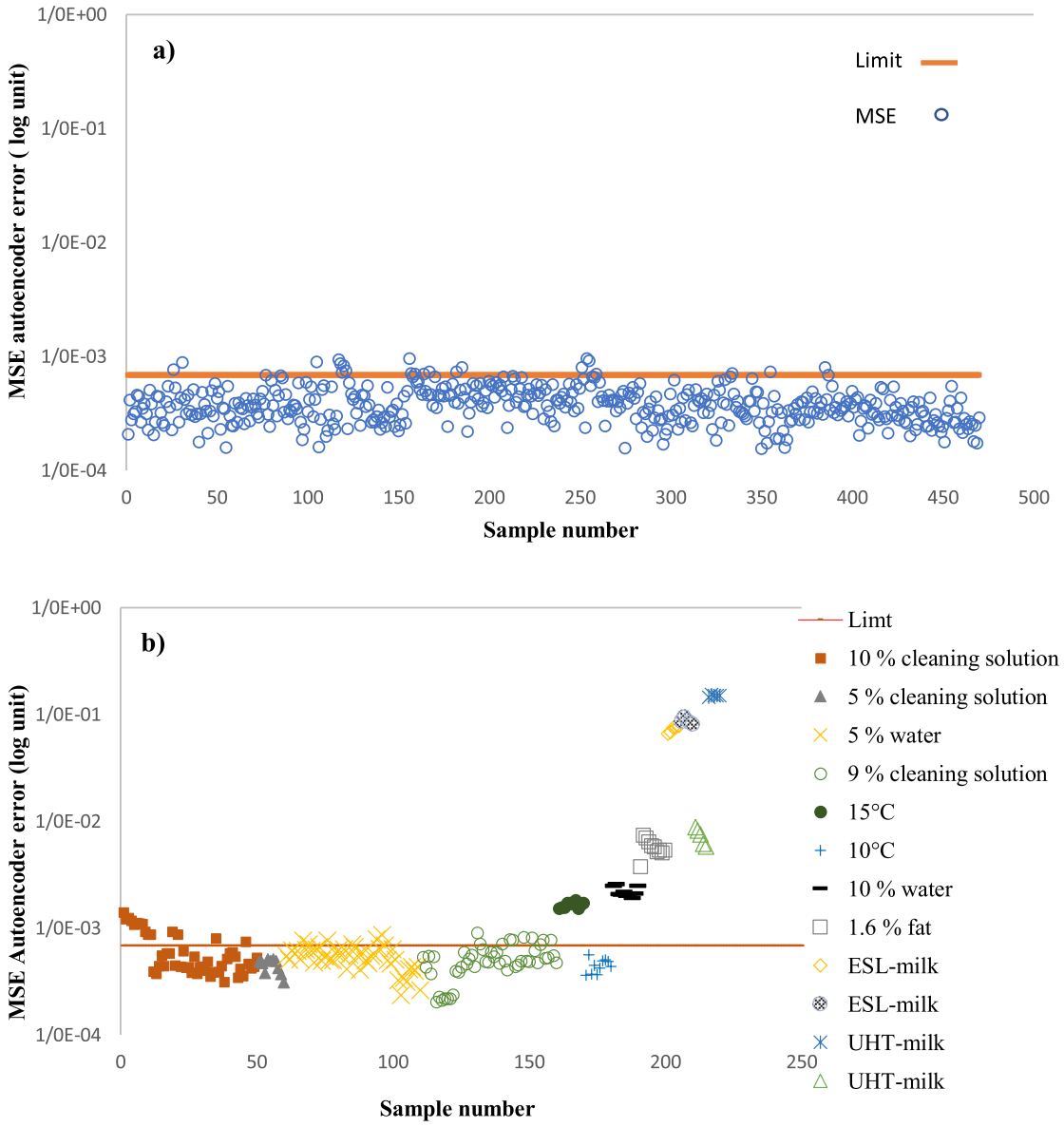


**Fig. 3.** Graph for training set (n = 470) showing the mean squared error of input and output of raw spectrum (MSE) at 5 °C (3.a) and graph for prediction set (n = 220) showing MSE of input and output of raw spectrum by using autoencoder (3.b). Limit is equal to upper range of confidential interval of training set with significance level of 5%. ESL-milks are from different suppliers.

the result.

Two groups of data sets (training set and prediction set) were created. For the training set, 470 spectra were taken from different packages of UHT-milk (ultra-high-temperature processing milk) with 1.5% of fat from the brand "Weihenstephan", Germany. For the prediction set, different types of samples were measured. As each company has the specific production methods, firstly various samples such as milk with 1.5% fat, both ESL-milk (extended shelf life) and UHT-milk from various suppliers were purchased from local stores and measured in order to check if there are any detectable differences from each other or not. Then, the same samples used for training set were measured after heating up to 10 °C or 15 °C to determine the sensitivity of samples to the temperature. A sample with 1.6% fat was prepared by mixing 95 mL of 1.5% fat milk and 5 mL of 3.5% fat milk from the brand of "Weihenstephan". It was prepared in order to be checked if it can be distinguished from 1.5% fat milk. For cleaning pilot plants, 1 mL of concentrated substance "Anti Germ clean A-N 30" is diluted with 99 mL water. This liquid can be used as the common cleaning solution of hygienic operation. Finally, different concentrations of water (0.05 mL/mL and 0.1 mL/mL) or cleaning solution (0.05 mL/mL, 0.09 mL/mL, and 0.1 mL/mL) were added to the 1.5% fat UHT-milk from "Weihenstephan" and were measured after cooling down to 5 °C. The purpose behind was to understand how much water and cleaning solution can be detected in the milk.

### 2.4. Data pre-processing

Spectral pre-processing techniques are required to remove any irrelevant information including noise, uncertainties, variability, interactions and unrecognized features (Lacroix et al., 1997). In this study, data were divided into three categories including raw data, first derivative, and combination of both. At first, raw data without any pre-processing was used. Next, first Savitzky-Golay derivative with second polynomial order and window size of 15 was applied. Finally,

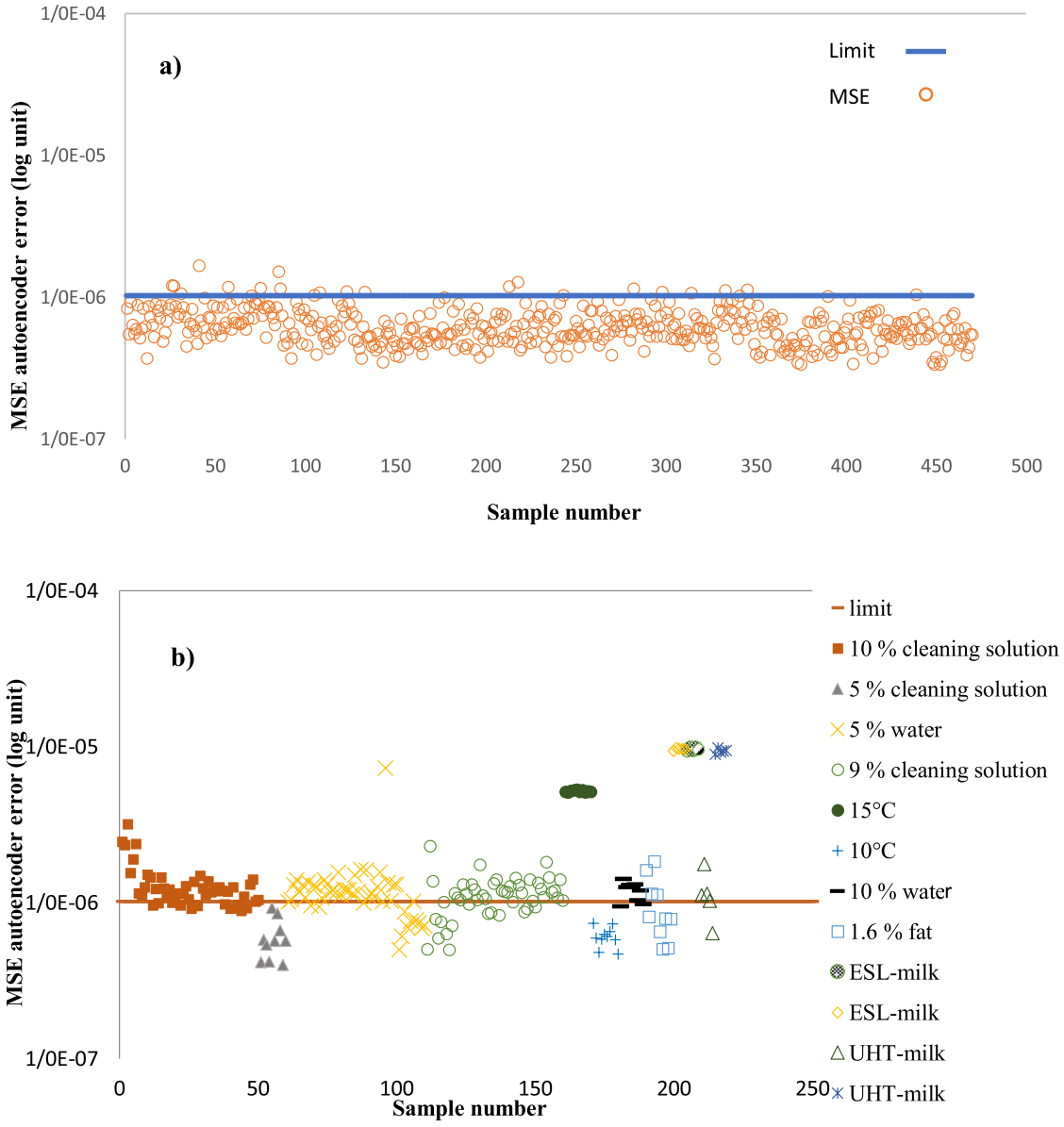


**Fig. 4.** Graph for training set (n = 470) showing the mean squared error of input and output of first derivative spectrum (MSE) at 5 °C (4.a) and graph for prediction set (n = 220) showing the mean squared error of input and output of first derivative spectrum (MSE) by using autoencoder (4.b). Limit is equal to upper range of confidential interval of training set with significance level of 5%. ESL-milks are from different suppliers.

each raw spectrum was supplemented with it's first derivative in order to investigate the effectiveness of such a combination. These methods were acquired by trials and checking different methods to show better and more accurate results. Other pre-processing methods such as smoothing, standard normal variate (SNV), multiplicative scatter correction (MSC), various normalizations and second derivative were applied on the spectral data but the results were not improved.

### 2.5. Autoencoder neural network

Autoencoder neural network is an unsupervised learning algorithm that can apply back-propagation, setting the output values to be equal to the inputs. In this feedforward network, the information moves in only one direction, forward, from the input nodes, through the hidden nodes to the output nodes (Fig. 1). There are no cycles or loops in the network.

Back-propagation is a widely used algorithm in training feedforward neural networks. The back-propagation neural network is a universal approximator given sufficient hidden units, multilayer feedforward sigmoidal network architectures can approximate virtually any function of interest to any desired degree of accuracy (Sharma et al., 2007). The maximum number of epochs is arbitrarily kept at 10000. An epoch describes the number of times the entire data set is processed by the algorithm. So, each time the algorithm has processed all samples in the dataset, an epoch has been completed. The tangent sigmoid transfer function (mathematically equivalent to tanh(x)) was used for hidden neurons, however, the identity transfer function was employed for the output neurons. The tangent sigmoid used here is given in the following equation:

$$f(x) = \frac{2}{1 + e^{-2x}} - 1 \quad (1)$$

Where x denotes weighted sum of the inputs.

As can be observed in Fig. 1, the input is compressed to the hidden layer(s) called encoding. Then, these compressed vector nodes are fully

**Table 2**
Percentage of autoencoder errors higher than limit of detection, and lower than it as well as mean and standard deviation for samples in the prediction set by using first derivative.

| Sample | Percentage of values lower than limit (%) | Percentage of values higher than limit (%) | Average of MSE | Standard deviation of MSE |
|---|---|---|---|---|
| 5% cleaning solution | 100 | 0 | 5.9E-07 | 1.7E-07 |
| 9% cleaning solution | 34 | 66 | 1.1E-06 | 3.2E-07 |
| 10% cleaning solution | 24 | 76 | 1.3E-06 | 4.4E-07 |
| 5% water | 30 | 70 | 1.2E-06 | 9.1E-07 |
| 10% water | 20 | 80 | 1.2E-06 | 1.5E-07 |
| 10 °C | 100 | 0 | 6.0E-07 | 8.9E-08 |
| 15 °C | 0 | 100 | 5.2E-06 | 7.9E-08 |
| 1.6% fat | 60 | 40 | 9.8E-07 | 4.5E-07 |
| ESL-milk | 0 | 100 | 9.6E-06 | 1.4E-07 |
| ESL-milk | 0 | 100 | 9.6E-06 | 1.3E-07 |
| UHT-milk | 40 | 60 | 5.4E-07 | 1.5E-07 |
| UHT-milk | 0 | 100 | 1.1E-06 | 4.0E-07 |

**Table 3**
Percentage of autoencoder errors higher than limit of detection, and lower than it as well as mean and standard deviation for samples in the prediction set by using combination of raw data and first derivative.

| Sample | Percentage of values lower than limit (%) | Percentage of values higher than limit (%) | Average of MSE | Standard deviation of MSE |
|---|---|---|---|---|
| 5% cleaning solution | 100 | 0 | 2.8E-04 | 3.5E-05 |
| 9% cleaning solution | 72 | 28 | 3.7E-04 | 1.5E-04 |
| 10% cleaning solution | 70 | 30 | 3.5E-04 | 1.5E-04 |
| 5% water | 62 | 38 | 3.7E-04 | 2.2E-04 |
| 10% water | 0 | 100 | 1.1E-03 | 1.1E-04 |
| 10 °C | 100 | 0 | 2.5E-04 | 4.3E-05 |
| 15 °C | 0 | 100 | 1.0E-03 | 8.9E-05 |
| 1.6% fat | 0 | 100 | 5.6E-03 | 1.7E-03 |
| ESL-milk | 0 | 100 | 5.8E-02 | 1.7E-03 |
| ESL-milk | 0 | 100 | 7.6E-02 | 3.4E-03 |
| UHT-milk | 0 | 100 | 5.8E-03 | 1.7E-03 |
| UHT-milk | 0 | 100 | 3.7E-02 | 1.0E-03 |

connected to a greater number of nodes in the hidden layer, until the number of nodes reaches the dimension of the input representation. This process is called decoding. In other words, autoencoder is trying to learn an approximation to the identity function, so as the output is similar to input.

In this work, a five-layer (I*10*5*10*O) feedforward back-propagation neural network was applied through experimental investigation of various internal parameters to detect changes in the milk processing. At the beginning, the autoencoder was trained with raw spectra and every 10th wavenumber was used for the input of autoencoder. In order to train the autoencoder, data was split to the training set and testing set. To avoid overfitting, training was stopped when the error of the test data increased. Training lasted 1–3 min based on the number of neurons. To analyse a spectrum, the sum of squared of differences between input and output of the autoencoder is calculated using the following equation:

$$\mathrm{MSE} = \left(\vec{x}_{out} - \vec{x}_{in}\right)\left(\vec{x}_{out} - \vec{x}_{in}\right)^{T} \Big/ n \quad (2)$$

MSE is the mean squared error of a spectrum and n is equal to the number of wavenumber channels used (441 is for every 10th wavenumber of the whole spectrum), $\vec{x}_{in}$ and $\vec{x}_{out}$ are the spectra (as a vector) for the input and output of the autoencoder, respectively. T indicated the transposed vector. A confidence interval was computed for training set by the formula mentioned below:

$$\text{Confidence Interval} = \overline{MSE} \pm t(\alpha, f) \times s \quad (3)$$

$\overline{MSE}$ is the average of autoencoder errors (MSE) of all training spectra, $t$ is t distribution, $f$ degree of freedom, and α significance level. In addition, s is equal to standard deviation of autoencoder errors.

220 spectra with the same wavenumbers as the training set were utilized for prediction set. In order to identify if the samples in the prediction set are different from training set, the upper limit of confidence interval of training set was used. A significance level of 5% was utilized. In the prediction set, the values higher than upper limit were considered as detectable values. For each sample, the measurement was done several times and after calculating MSE for each replication, the average and standard deviation were computed. The same procedure was applied for first derivative and the combination spectra. However, the main difference was the number of inputs in the neural network. When all the wavenumbers of transformed first derivative spectra were used as input (4407 wavenumbers), every 10th wavenumbers were utilized for combination spectra. The MATLAB software's proprietary script language (version of 2019b) with Deep Learning Toolbox (version of 13.0) have been used for the purpose of developing the autoencoder neural network.

## 3. Results and discussion

Fig. 2 summarize NIR absorption intensities related to the characteristics of the chemical bonds within milk ingredients.

The graph is showing raw data without any pre-processing with different peaks that depict specific components of milk, e.g. –CH groups within the chains of fatty acid molecules, carbonyl groups in ester linkages of fat molecules, peptide linkages between amino acids of protein molecules, and –OH groups in lactose and water. NIR spectroscopic technique relies on different energy states using excitations of higher quanta transitions, i.e. first overtones and binary combinations of fundamental molecular vibrations. NIR spectra contain information about molecular vibration of certain groups and harmonics of said vibrations. The stretching vibration and stretching–bending combinations of various chemical groups of chemical bonding (O–H, N–H, C–H, S–H) in the molecules have natural vibrational frequencies, of which overtones and combination tones appear in the region from 9000 to 4000 $cm^{-1}$. Proteins are mainly composed of amino acid molecules through peptide bonds (–CO–NH–), and their N–H bond is the main absorption group. The first-order frequency doubling of the N–H bond stretching vibration would appear in specific wavenumber while the frequency of combination of the stretching vibration and bending vibration would be at a different wavenumber so several peaks related to protein are shown. The spectral features of some components are mainly distributed throughout some spectral regions whereas others would appear in the specific wavenumber. Due to the coupling of fundamental frequency, double frequency and combined frequency, polyatomic molecules have many absorption bands in the whole near infrared region, and it is difficult to accurately distinguish the attribution of near infrared bands (Chang et al., 2020). N–H, O–H and C–H bonds of NIR are located around 8500–4000 $cm^{-1}$. In the range of 9000–4000 $cm^{-1}$, the NIR spectra showed nine main features, whose representative chemical groups have been marked in Fig. 2. These results show that compared with the range of 12000–9000 $cm^{-1}$, more useful information can be

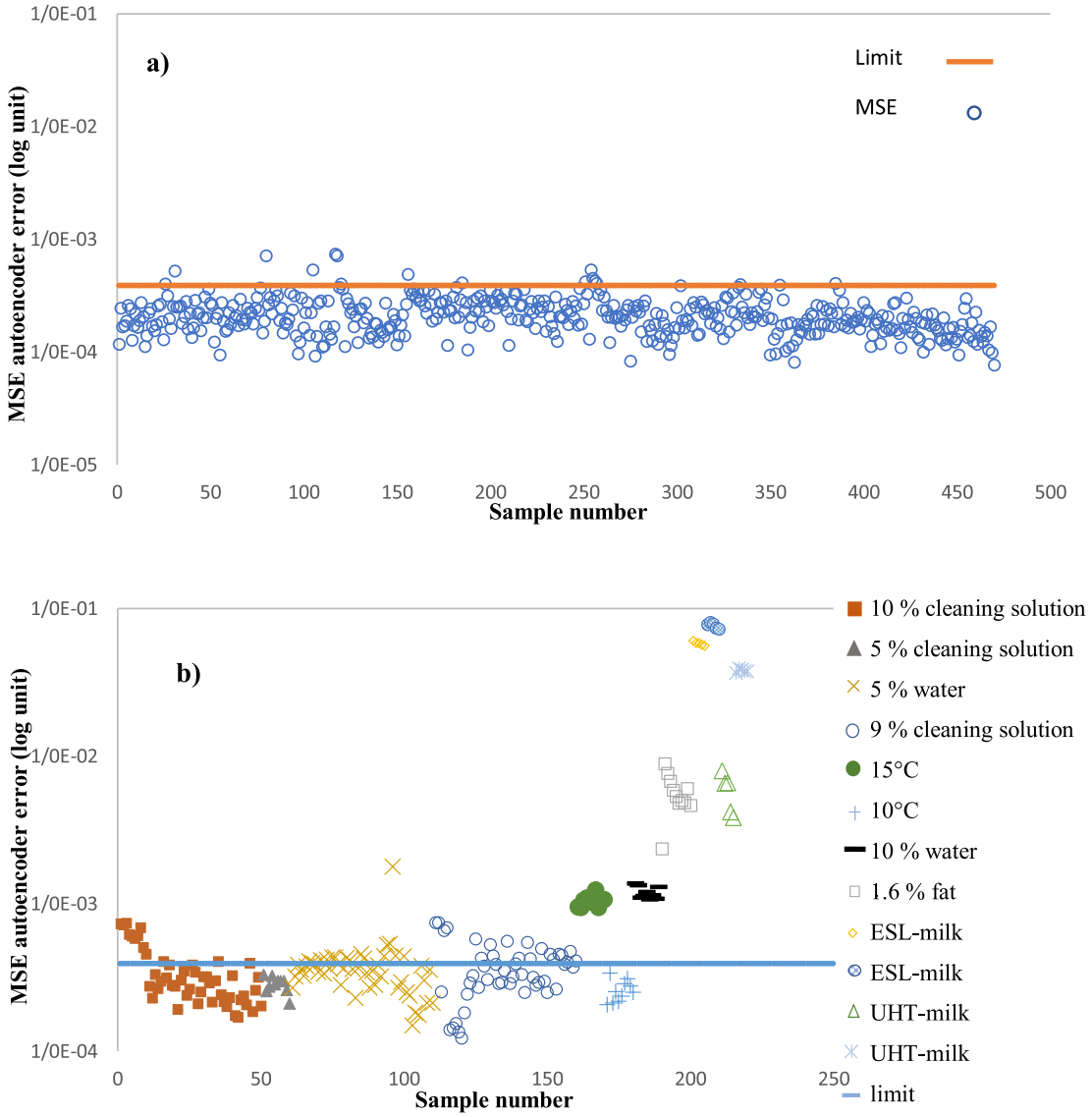


**Fig. 5.** Graph for training set (n = 470) showing the mean squared error of input and output of pre-processed spectrum (combination of raw and first derivative) at 5 °C (5.a) and graph for prediction set (n = 220) showing the mean squared error of input and output of pre-processed spectrum (combination of raw and first derivative) by using autoencoder in the cow milk (5.b). Limit is equal to upper range of confidential interval of training set with significance level of 5%. ESL-milks are from different suppliers.

acquired from the range of 9000–4000 $cm^{-1}$. As can be seen, a prominent and broad band around 6900 $cm^{-1}$ (the first overtone of O–H stretching vibration) has been attributed to water. In addition, water shows strong absorption bands of O–H stretching and bending at around 5100–5300 $cm^{-1}$ in NIR spectroscopy. The band at 5000 $cm^{-1}$ has been associated with protein as amide, besides absorption bands at about 6250 $cm^{-1}$ (the first harmonic free N–H stretch/amide II combination). Furthermore, other bands corresponding to fat content appeared at 5800 $cm^{-1}$ (the first overtone of –CH stretching) (dos Santos Pereira, 2020). Milk contains about 88% water, which produces strong bands in NIR around 10400, 6950, 5150, and 4820 $cm^{-1}$, which overlap with some bands of interest creating noise (Tsenkova et al., 1999). In NIR, for instance, characteristic absorption bands of milk components such as fat may be affected by the high absorption by water in combination with the strong light scattering by the fat globules (Aernouts et al., 2011).

### 3.1. Detection of changes by the autoencoder using raw data

At the first level, raw spectra without any pre-processing were used. Every 10th wavenumber was utilized as the input of this neural network. The average and standard deviation of MSE for the training set was calculated (3.9 ± 1.4) E−04. The upper limit of the confidence interval with significance level of 5%, mentioned as limit of detection is equal to 6.8E-04. Generally, the autoencoder errors which are lower than this limit are not considered as abnormal whereas higher values can show the detection capability of this neural network. Table 1 presents results of the prediction set including percentage of autoencoder errors which are higher and lower compared to the limit, as well as average and standard deviation. Fig. 3.a illustrates the autoencoder errors of the training set at 5 °C by using raw data without any pre-processing. The graph in Fig. 3.b shows the individual autoencoder errors for the prediction set as well as the limit of detection.

The examination of the prediction set reveals that all the replications related to milk with 5% cleaning solution are lower than 6.8E-04 which

is specified as limit. It means that 5% cleaning solution cannot be detected by the autoencoder. However, in milk with 9% cleaning solution, 80% of replications are lower than the limit. The average of replications shows lower value than the limit, it points that 9% of cleaning solution in milk is still undetectable. By having a closer look to the samples with 10% cleaning solution, it can be observed that 30% of values are higher than the limit and the average of replications is 6.2E-04 which is very close to the limit. According to Fig. 3.b, the values of 10% cleaning solution spread around the limit clearly. The same procedure for 5% water confirms that just 14% of repetitions are higher than the limit and also 5.4E-04 and 1.3E-04 are mentioned as average and standard deviation respectively. Based on the results in Table 1, the average of this sample is lower than 6.8E-04, so 5% water is not detectable by using autoencoder. However, by testing further concentration of water (10%), all the repetitions are above the limit. As a result, it is concluded that 10% water in milk is recognised by this neural network. As the 1.5% fat milk was used for training the autoencoder, 1.6% fat milk was tested in order to investigate if other concentrations of fat are detectable or not. According to Fig. 3.b, all the values related to this sample are clearly above the limit which demonstrates the fact that the autoencoder can distinguish at least the difference of 0.1% fat in the milk. Temperature variation was another variable which was checked by this neural network. While the training set was measured at 5 °C, neural network behaves differently with the samples measured at the temperature of 10 °C and 15 °C. Based on Table 1, 100% of autoencoder errors of samples measured at 10 °C are lower than 6.8E-04. In contrast, all the samples measured at 15 °C are above the limit and can be distinguished. The reason is sensitivity of NIR to the temperature as it works based on vibrational movements. In the further steps, when two ESL-milk with 1.5% fat from various suppliers are used in prediction set, the autoencoder errors of them are intensively high. It depicts to the different procedures between ESL-milk and UHT-milk preparation which can be easily detect by the autoencoder. In addition, according to Fig. 3.b, all samples with the same characteristics to the training sample from several suppliers are higher than the limit which points to the specific production methods of each companies.

Generally, by implementing the special procedure of using raw data, three hidden layers (10*5*10) and 441 inputs (every 10th wavenumber), some changes in the milk like adding 10% water, 0.1% fat, 10 °C difference in temperature of samples, and various production methods are detectable. However, milks with 10% cleaning solution is overlapped with the limit of detection. The technique is rapid, nondestructive, precise and cost-effective, compared with other laboratory techniques. Kasemsumran et al. (2007) confirmed the feasibility of NIR to detect and to quantify water or whey adulterants in bovine milk. The study concluded that, for the detection of whey or water adulterant, pretreated spectra with MSC and 2nd derivative method were needed. Jaiswal et al. (2017) reported that Fourier Transform Infrared spectroscopy is a rapid method for detection and quantification of anionic detergent in milk. For determination of detergent in milk, partial least squares regression was used. Indicating the potential of this method without much sample preparation and data treatment was promising. Despite of other works, using NIR and autoencoder can investigate and detect several anomalies simultaneously which consider as the advantage of this procedure. These changes are not always small enough so that the regulations in Germany are fulfilled. For instance, it is not allowed to add water to milk. For semi-skimmed milk (heat-treated low-fat milk), where a minimum of 1.5% and a maximum of 1.8% fat is required, the autoencoder can provide useful information. However, the detection of 10% of detergent is not enough and must be improved.

### 3.2. Detection of changes by the autoencoder using first derivative data

To investigate the effects of pre-processing on the improvement of detection, the first derivative with a window size of 15 was applied and all the wavenumbers of the pre-processed spectra were utilized as the input of the autoencoder neural network. In this case, the limit of detection which is calculated by using confidence interval with the significance level of 5 percent is equal to 1.0E-06. Also, the average and standard deviation of MSE for training set are 6.4E-07 and 1.9E-07 respectively. Fig. 4 displays the autoencoder errors of training set and prediction set by using pre-processed data.

Table 2 states the results of prediction set which are pre-processed by first derivative Savitzky-Golay filter.

According to Fig. 4.b all the replications of milk with 5% cleaning solution are lower than the limit which illustrates that it is not detectable, even by using derivative data. However, confirming to Table 2, for the samples with 9% and 10% cleaning solution, the percentage of values higher than the limit are 66% and 76% respectively. These two samples show the average of 1.1E-06 and 1.3E-06, which both are higher than the limit. Therefore, it would signify the capability of autoencoder to detect the 9% cleaning solution in the milk. The average of 1.2E-06 (higher than the limit of detection) for milk with 5% water showing the usefulness of autoencoder to determine this concentration of water inside the milk. Fig. 4.b illustrates that despite of samples were measured at 10 °C which the autoencoder cannot distinguish them from the training set, samples measured at 15 °C can be properly detected. Also, by looking closer at Fig. 4.b, it is obvious that autoencoder errors of milk with 1.6% fat are spreading around the limit. Additionally, the average of 9.8E-07 for this sample shows that it can be overlapped with the limit of detection so the difference of 0.1 percent fat can't be recognised reliably. The results of UHT and ESL milk from several suppliers demonstrate that autoencoder is able to detect some changes in these samples in compare with the training set.

In order to compare the usage of raw data to the first derivative in this procedure, it can be concluded that using first derivative with window size of 15 would be more effective to detect the water and cleaning solution in the milk. In contrast, to recognize the difference in the fat concentration of milk, the results of raw data show much more accurate prediction.

### 3.3. Detection of changes by the autoencoder using combination of raw and first derivative spectrum

In pursuance of finding better results, the raw spectra were combined with their first derivative, and every 10th wavenumbers were used as the input of the autoencoder. The average of autoencoder errors for training set is equal to 2.1E-04 ± 8.6E-05 and the limit of detection is 3.8E-04. Table 3 shows the percentage of autoencoder errors higher and lower than the limit of detection as well as mean and standard deviation for samples in the prediction set. SNV (standard normal variate) was applied after combining the raw and derivative spectra but it did not improve the results. Fig. 5 presents the graph of training and prediction set which are prepared by autoencoder from the combination of raw and first derivative spectrum.

By using the mentioned procedure, all replications related to the milk with 5% cleaning solution are lower than 3.8E-04 considered as the limit of detection. However, the percentage of autoencoder errors above the limit for 9% and 10% cleaning solution are 28% and 30% respectively. As a result, the lower average of samples with 9% and 10% cleaning solution (compared to the limit of detection) imply the lower efficiency of combination spectrum for cleaning solution detection. However, these averages are very close to the limit of detection. The average of mean squared errors for milk with 5% water is 3.7E-04 which is very close to the limit but still lower. In contrast, samples with 10% water can be detected properly. By taking a closer look to Fig. 4.b, despite of samples measured at 15 °C, all the milks measured at 10 °C are not detectable. 0.1% fat can be recognised easily by using combination of raw and first derivative spectra. In addition, various production methods can be distinguished from each other. The Fourier Transform Infrared spectroscopy was before evaluated as a promising method for detection of anionic detergents in milk. Principal component analysis showed

discrete clustering of samples based on level of detergents in milk (Jaiswal et al., 2017). Other work was employed on NIR spectroscopy and PLS algorithms for the identification and quantification of goat milk adulteration by adding cow milk, besides the determination of their fat and protein contents. Therefore, the proposed methodology proved to be a useful, fast and non-destructive tool for screening the quality of goat milk in terms of its adulteration with cow milk, in addition to the quantification of its fat and protein contents (dos Santos Pereira, 2020).

Using the combination of raw and first derivative spectra can improve the autoencoders detection abilities compared to raw data alone. However, for detecting the cleaning solution and water inside the milk, using only first derivative alone would be more effective. On the other hand, the combination of the first derivative with the raw unprocessed signal was useful to detect fat, temperature, and production method changes. Overall, as this combination can detect more anomalies, it is recommended. In addition, pre-processed spectra usually are better than raw data to detect anomalies which can be consider as advantage of this method but pre-processing in general could be time consuming or induce more cost. This procedure can detect several anomalies simultaneously so it is superior to existing methods.

As a result, the detection of anomalies during milk processing is an important sector as it effects the quality and safety of the final product. As a new method of unsupervised machine learning, deep autoencoder represent a powerful capability in anomaly detection. By this method most issues can be detected early in the process and potentially be corrected. As a spectroscopic method, it is rapid with minimal sample preparation which is the most important advantage. In this contribution, the potential of an autoencoder for detecting possible changes during milk processing was evaluated. The performance of NIR spectroscopy was investigated as well. Raw spectra acquired by NIR spectrometer were compared with their first derivative and combination of both in order to find the best anomaly detection method. It can be concluded that most abnormal states in milk processing can be detected by this method. Using first derivative was helpful to detect 5% added water and 9% cleaning solution in the milk, however, a combination of raw spectra with their first derivative was more effective to detect the difference of 0.1% fat. Both procedures were able to detect 10 °C temperature variation, as well as variations in various other production parameters. In summary, this innovative method shows great promise for use during milk processing and beyond to detect anomalies during the production. It can be adapted easily for new processes, because it is a self-learning unsupervised evaluation method. Therefore, this approach can be adopted for the quality control of pretty much all liquid beverages such as juice, lemonade and even further liquid products like oil.

## Credit statement

Pegah Sadeghi Vasafi: Methodology, Software, Formal analysis, validation, Investigation, Writing-Original Draft, Writing-Review and Editing. Olivier Paquet-Durand: Software, Writing-Review and Editing. Kim Brettschneider: Methodology, Resources, Writing-Review and Editing. Jörg Hinrichs: Supervision, Writing-Review and Editing. Bernd Hitzmann: Conceptualization, Project administration, Supervision, Writing-Review and Editing.

## Acknowledgments

This IGF project of the FEI is/was supported via AiF (20200N) within the programme for promoting the Industrial Collective Research (IGF) of the German Ministry of Economics and Energy (BMWi), bases on a resolution of the German Parliament for which the authors are grateful.

## References

Aernouts, B., Polshin, E., Saeys, W., Lammertyn, J., 2011. Mid-infrared spectrometry of milk for dairy metabolomics: a comparison of two sampling techniques and effect of homogenization. Anal. Chim. Acta 705 (1–2), 88–97.

Balabin, R.M., Smirnov, S.V., 2011. Melamine detection by mid- and near-infrared (MIR/NIR) spectroscopy: a quick and sensitive method for dairy products analysis including liquid milk, infant formula, and milk powder. Talanta 85 (1), 562–568. https://doi.org/10.1016/j.talanta.2011.04.026.

Bylund, G., 1995. Dairy Processing Handbook, *Tetra Pak Processing Systems*, pp. 65–385.

Chandan, R.C., 2011. Dairy ingredients for food processing: an overview. Dairy ingredients for food processing 3–33.

Chang, S., Yin, C., Liang, S., Lu, M., Wang, P., Li, Z., 2020. Confirmation of brand identification in infant formulas by using near-infrared spectroscopy fingerprints. Analytical Methods 12 (19), 2469–2475.

Curto, B., Moreno, V., García-Esteban, J.A., Blanco, F.J., González, I., Vivar, A., Revilla, I., 2020. Accurate prediction of sensory attributes of cheese using near-infrared spectroscopy based on artificial neural network. Sensors 20 (12), 3566.

dos Santos Pereira, E.V., de Sousa Fernandes, D.D., de Araújo, M.C.U., Diniz, P.H.G.D., Maciel, M.I.S., 2020. Simultaneous Determination of Goat Milk Adulteration with Cow Milk and Their Fat and Protein Contents Using NIR Spectroscopy and PLS Algorithms. LWT, p. 109427.

Eerikäinen, T., Linko, P., Linko, S., Siimes, T., Zhu, Y.H., 1993. Fuzzy logic and neural network applications in food science and technology. Trends Food Sci. Technol. 4 (8), 237–242. https://doi.org/10.1016/0924-2244(93)90137-Y.

Fu, X., Kim, M.S., Chao, K., Qin, J., Lim, J., Lee, H., Garrido-Varo, A., Pérez-Marín, D., Ying, Y., 2014. Detection of melamine in milk powders based on NIR hyperspectral imaging and spectral similarity analyses. J. Food Eng. 124, 97–104. https://doi.org/10.1016/j.jfoodeng.2013.09.023.

Gong, D., Liu, L., Le, V., Saha, B., Mansour, M.R., Venkatesh, S., Van Den Hengel, A., 2019. Memorizing Normality to Detect Anomaly: Memory-Augmented Deep Autoencoder for Unsupervised Anomaly Detection. Presented at the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, Seoul, Korea (South), pp. 1705–1714.

Heald, C.W., Kim, T., Sischo, W.M., Cooper, J.B., Wolfgang, D.R., 2000. A computerized mastitis decision and using farm-based records: an artificial neural network approach. J. Dairy Sci. 83 (4), 711–720. https://doi.org/10.3168/jds.S0022-0302(00)74933-2.

Hillerton, J.E., Berry, E.A., 2004. Quality of the milk supply: European regulations versus practice. NMC Annual Meeting Proceedings 207, 214.

Hitzmann, B., Hauselmann, R., Niemoeller, A., Sangi, D., Traenkle, J., Glassey, J., 2015. Process analytical technologies in food industry–challenges and benefits: a status report and recommendations. Biotechnol. J. 10 (8), 1095–1100.

Huang, H., Yu, H., Xu, H., Ying, Y., 2008. Near-infrared spectroscopy for on/in-line monitoring of quality in foods and beverages: a review. J. Food Eng. 87 (3), 303–313. https://doi.org/10.1016/j.jfoodeng.2007.12.022.

Huang, M., Kim, M.S., Delwiche, S.R., Chao, K., Qin, J., Mo, C., Esquerre, C., Zhu, Q., 2016. Quantitative analysis of melamine in milk powders using near-infrared hyperspectral imaging and band ratio. J. Food Eng. 181, 10–19. https://doi.org/10.1016/j.jfoodeng.2016.02.017.

Jaiswal, P., Jha, S.N., Kaur, J., Borah, A., 2017. Detection and quantification of anionic detergent (lissapol) in milk using attenuated total reflectance-Fourier Transform Infrared spectroscopy. Food Chem. 221, 815–821.

Kasemsumran, S., Thanapase, W., Kiatsoonthon, A., 2007. Feasibility of near-infrared spectroscopy to detect and to quantify adulterants in cow milk. Anal. Sci. 23 (7), 907–910.

Kawasaki, M., Kawamura, S., Tsukahara, M., Morita, S., Komiya, M., Natsuga, M., 2008. Near-infrared spectroscopic sensing system for on-line milk quality assessment in a milking robot. Comput. Electron. Agric. 63 (1), 22–27. https://doi.org/10.1016/j.compag.2008.01.006.

Khan, A., Munir, M.T., Yu, W., Young, B.R., 2020. Near-infrared spectroscopy and data analysis for predicting milk powder quality attributes. International Journal of Dairy Technology. https://doi.org/10.1111/1471-0307.12734.

Lacroix, R., Salehi, F., Yang, X.Z., Wade, K.M., 1997. Effects of data pre-processing on the performance of artificial neural network for dairy yield prediction and cow culling classification. Transactions of the ASAE 40, 839–846. American Society of Agricultural Engineers.

Lu, C., Xiang, B., Hao, G., Xu, J., Wang, Z., Chen, C., 2009. Rapid detection of melamine in milk powder by near-infrared spectroscopy. Journal of Near-Infrared Spectroscopy 17, 59–67. https://doi.org/10.1255/jnirs.829.

Mauer, L.J., Chernyshova, A.A., Hiatt, A., Deering, A., Davis, R., 2009. Melamine detection in infant formula powder using near- and mid-infrared spectroscopy. J. Agric. Food Chem. 57, 3974–3980. https://doi.org/10.1021/jf900587m.

Mittal, G.S., Zhang, J., 2000. Use of artificial neural network to predict temperature, moisture, and fat in slab-shaped foods with edible coatings during deep-fat frying. J. Food Sci. 65 (6), 978–983. https://doi.org/10.1111/j.1365-2621.2000.tb09403.x.

Núñez-Sánchez, N., Martínez-Marín, A.L., Polvillo, O., Fernández-Cabanás, V.M., Carrizosa, J., Urrutia, B., Serradilla, J.M., 2016. Near Infrared Spectroscopy (NIRS) for the determination of the milk fat fatty acid profile of goats. Food Chem. 190, 244–252. https://doi.org/10.1016/j.foodchem.2015.05.083.

Paquet, J., Lacroix, C., Thibault, J., 2000. Modelling of pH and acidity for individual cheese production. J. Dairy Sci. 83 (11), 2393–2409. https://doi.org/10.3168/jds.S0022-0302(00)75129-0.

Sanzogni, L., Kerr, D., 2001. Milk production estimates using feed forward artificial neural networks. Comput. Electron. Agric. 32 (1), 21–30.

Scott, D.M., McCann, H., 2005. Process Imaging for Automatic Control. CRC Press Taylor and Francis Group, pp. 85–125. https://doi.org/10.1201/9781420028195.

Sharifi, M., Young, B., 2011. 3-Dimensional spatial monitoring of tanks for the milk processing industry using electrical resistance tomography. J. Food Eng. 105 (2), 312–319. https://doi.org/10.1016/j.jfoodeng.2011.02.041.

Sharma, A.K., Sharma, R.K., Kasana, H.S., 2007. Prediction of first lactation 305-day milk yield in Karan Fries dairy cattle using ANN modeling. Appl. Soft Comput. 7 (3), 1112–1120. https://doi.org/10.1016/j.asoc.2006.07.002.

Tsenkova, R., Atanassova, S., Toyoda, K., Ozaki, Y., Itoh, K., Fearn, T., 1999. Near-infrared spectroscopy for dairy management: measurement of unhomogenized milk composition. J. Dairy Sci. 82 (11), 2344–2351. https://doi.org/10.3168/jds.S0022-0302(99)75484-6.

Widyanto, M.R., Nobuhara, H., Kawamoto, K., Hirota, K., Kusumoputro, B., 2005. Improving recognition and generalization capability of back-propagation NN using a self-organized network inspired by immune algorithm (SONIA). Appl. Soft Comput. 6 (1), 72–84. https://doi.org/10.1016/j.asoc.2004.10.008.

Williams, R.A., Beck, M.S., 1995. Process Tomography; Principles, Techniques and Applications. Butterworth-Heinemann Ltd., pp. 20–100

Workman, J., Koch, M., Veltkamp, D.J., 2003. Process analytical chemistry. Anal. Chem. 75 (12), 2859–2876. https://doi.org/10.1021/ac0301393.

Zettel, V., Ahmad, M.H., Beltramo, T., Hermannseder, B., Hitzemann, A., Nache, M., Paquet-Durand, O., Schöck, T., Hecker, F., Hitzmann, B., 2016. Supervision of food manufacturing processes using optical process analyzers–an overview. ChemBioEng Reviews 3 (5), 219–228.

## 3.2. Establishing a novel procedure to detect deviations from standard milk processing by using online Raman spectroscopy

By Pegah Sadeghi Vasafi, Jörg Hinrichs, Bernd Hitzmann. Published in Food Control, volume 131, page 108442. January 2022.

Food Control 131 (2022) 108442

Contents lists available at ScienceDirect

Food Control

journal homepage: www.elsevier.com/locate/foodcont

# Establishing a novel procedure to detect deviations from standard milk processing by using online Raman spectroscopy

Pegah Sadeghi Vasafi *, Jörg Hinrichs, Bernd Hitzmann

*Institute of Food Science and Biotechnology, University of Hohenheim, Garbenstr. 23, 70599, Stuttgart, Germany*

ARTICLE INFO

*Keywords:*
Autoencoder
Milk processing
Deviations detection
Gaussian process regression
Raman spectroscopy

ABSTRACT

Controlling milk processing steps is a crucial task as it affects the quality and safety of the final product. Using Raman spectrometer in combination with various evaluation techniques such as principal component analysis and regression, Gaussian process regression, and the autoencoder were checked to define an accurate method for detection of deviations from standard procedures. For this purpose, milk with 5% fat measured at 10 °C was considered as the reference milk. A temperature-controlled flow cell was used in a by-pass for online measurements. While the principal component regression was not able to predict the deviations, results demonstrate the capability of Gaussian process regression and the autoencoder to detect 5% added water and cleaning solution, 0.1% difference in fat content and variation of 5 °C in measurement temperature. It can be concluded that both procedures display promising results, however, the autoencoder can be trained once and used immediately for online supervision. Therefore, changes can be detected promptly, enabling companies to react instantly.

## 1. Introduction

Milk is a complete nutrient source for humans, so its quality and safety would be critical not only for producers but also for consumers. Thereby, to ensure the quality and safety of milk a prognostic tool based on online measurement data would be beneficial for fast detection of deviations from the standard product. Especially when the conventional methods do only analyze statistically selected samples are time-consuming, laborious, and require complicated preparation levels. During milk production, the sources of raw milk, ways of transportation and storage as well as deviations in processing parameters or cleaning can affect the quality and safety of final products. Therefore, there are some demands for time-resaved techniques to evaluate and control each step or in summary the final product until filling and packing (He et al., 2019).

Raman spectroscopy, based on an inelastic scattering effect, which can be observed when electromagnetic radiation interacts with matter, was first published by Raman and Krishnan (1928). Dijkstra et al. (2005) mentioned that there are various forms of vibration in molecules but Raman signals can be acquired by only a few of them. Raman signals can therefore be observed while the polarizability of the molecule changes. Raman spectroscopy has been formerly used in the dairy industry (Blanpain-Avet et al., 2012). To detect the artificial additives in milk, Rajapandiyan et al. (2015) illustrated a method for rapid determination of melamine in milk samples by surface-enhanced Raman spectroscopy technique. El-Abassy et al. (2011) illustrated the capability of visible Raman spectroscopy combined with partial least square regression (PLS) as an accurate and fast method for direct detection of milk fat. Also, Bernuy et al. (2008) explained the potential capacity of Fourier Transform (FT) Raman spectroscopy in the detection of total conjugated linoleic acid (CLA) in milk in 57 anhydrous milk fat samples. Another research was carried out by Moros et al. (2007) using Fourier Transform spectroscopy to identify fat content in milk powder. In their study, raw data without any pre-processing were employed. A partial least squares regression modelling with 8% mean relative prediction error was obtained. Stefanov et al. (2011) verified the feasibility of Fourier Transform Raman spectroscopy in combination with partial least square regression for the determination of individual or grouped *trans*-monounsaturated fatty acids (FAs) and conjugated linoleic acid in milk at two temperature conditions. In addition, Li et al. (2015) employed crystal violet as an internal standard and employed Raman spectroscopy with a 785 nm laser light to quantitative detect lactose in milk. The results demonstrated a proper linear relationship between lactose and peak intensity with $R^2$ of 0.99. More recently, Vaskova et al. (2016) compared phenylalanine and crystal violet for normalizing Raman identification and quantification of lactose. Furthermore, Acar-Soykut

* Corresponding author.
*E-mail address:* Pegah.sadeghivasafi@uni-hohenheim.de (P.S. Vasafi).

https://doi.org/10.1016/j.foodcont.2021.108442
Received 14 May 2021; Received in revised form 22 June 2021; Accepted 19 July 2021
Available online 21 July 2021

et al. (2018) demonstrated that Raman spectroscopy combined with principal component analysis (PCA) analysis can be applied to determine antibiotics and phages in raw milk samples.

Deep autoencoder is a powerful tool to model high-dimensional data in the unsupervised setting with the aim of detecting deviations from the standard. In details, it can learn a "normal" profile (spectra from reference sample), thereby define the samples which are not conforming to the normal profile (standard products with no quality or safety issues) by producing higher reconstruction error than the reference samples. The higher autoencoder error can be used as a criterion for the detection of anomalies (Gong et al., 2019). The autoencoder has not been extensively applied in food science in particular dairy technology. Vasafi et al. (2021) used the autoencoder for anomaly detection during milk processing based on near-infrared spectroscopy. Here the prediction capability of this evaluation method performed best.

Gaussian process regression (GPR) can be applied to solve regression problems and requires only a few samples for modelling (Wang & Chaib-draa, 2016). Wang et al. (2021) have built a soft sensor model based on Gaussian process regression in order to monitor the chlortetracycline fermentation process based on online measurable parameters and the total sugar content of the fermentation broth. However, this procedure has not been used intensively in food science. Recently, a novel prediction model of diet nutrients digestibility of dairy cows based on Gaussian process regression was developed (Fu et al., 2019). Also, Liu et al. (2019) proposed a non-destructive detection method for packaged food based on the generalized Gaussian model for Raman scattering images. In this study, line-scan Raman scattering images were used. The Raman peaks of the scattering image were extracted, and the attenuation information of the peaks far from the laser point was imported into the established generalized Gaussian model. Analysis of the histogram of residual distribution revealed that the difference in residual distribution was enhanced, and an appropriate threshold was selected to separate the Raman baseline correction spectrum of the internal materials. Generalized Gaussian model was able to separate food and package Raman peaks.

The aim of this research was to develop an appropriate technique for detecting deviations from standard samples during the milk processing (such as changes in fat, temperature, added water or contamination of cleaning solution) by using Raman spectroscopy. A flow cell was established in a by-pass, then the spectra were collected online. Principal component regression, Gaussian process regression as well as an autoencoder were tested to be valuable to identify such deviations from the standard by means of spectra comparison.

## 2. Material and method

### 2.1. Material

In this contribution, two different types of homogenized milk with 1.5% and 3.5% fat content were used from the "Weihenstephan" brand, Germany. Milk with a concentration of 1.5% fat was considered as the reference sample (standard). Samples were kept at the temperature of 5 °C before opening the packages. "Anti-Germ clean A-N 30" was used as a test cleaning solution to prepare contaminated samples.

The spectra evaluation was carried out on a Windows PC 10. As software for the principal component analysis and regression, the Unscrambler (version of 10.3) as well as for all other evaluation techniques, MATLAB (version of 2019b) with the Deep Learning Toolbox version 13.0 were applied.

### 2.2. Sample preparation

As the reference sample UHT milk (ultra-high-temperature processing milk) with 1.5% of fat was used. Then the reference sample was modified by some changes. A sample with 1.6% fat was prepared by mixing 95 mL of 1.5% fat milk and 5 mL of 3.5% fat milk. The same procedure was done to create concentrations of 1.7% and 1.8% fat by mixing various amounts. So, the ability of different evaluation techniques to distinguish different fat contents in milk was checked. For cleaning plants, 1 mL of concentrated substance "Anti Germ clean A-N 30" is diluted with 99 mL water. This liquid is used as the common cleaning solution of hygienic operation. Subsequently, various concentrations of water or cleaning solution (0.05 L/L and 0.1 L/L) were added to the reference milk. The purpose behind it was to understand how much water and cleaning solution can be detected by using different evaluation procedures. All the modified samples were measured after reaching the temperature of 10 °C. For the final step reference milk was measured after warming up to 15 °C or 20 °C to determine the temperature sensitivity of the analysis.

### 2.3. Raman spectroscopy

Raman measurements were carried out using an INNO-SPEC Raman 785 Spectrometer (INNO-SPEC GmbH, Germany) with a laser excitation wavelength of 784.98 nm. The signal was collected on a thermoelectric cooled back-thinned CCD (charge-coupled device) sensor. All Raman spectra consisted of a scan at 1 $cm^{-1}$ resolution across the spectral range of 65–3290 $cm^{-1}$. Integration time (IT) of 20 s and power of less than 500 mW were employed.

For spectra acquisition, a high-quality quartz flow cell of 1 mm path length including temperature controlling option for liquids was implemented (Hellma Analytics, Germany). The volume of the flow cell was less than 1 mL. The flow cell was placed 12 mm far away from the outlet of a temperature-stabilized laser to ensure proper focusing on the sample. In order to adjust the cold temperature of samples during the measurement, a water bath with a temperature of 5 °C was utilized. A beaker holding milk at 5 °C was left in the water bath and connected to the flow cell by plastic tubes. In order to keep the temperature stable and simulate online process measurement conditions in a by-pass, milk was circulated through the flow cell during the measurement. Right after starting the milk circulation, its temperature increased to 10 °C because of the heat produced by the laser as well as the length of plastic tubes, which can be affected by room temperature. Spectra were acquired by applying a continuous mode of measurement during the milk circulation through the flow cell. 139 spectra were acquired from three different packages of reference milk and 213 spectra were obtained from nine modified samples including several repetitions for each sample.

### 2.4. Pre-processing algorithms

Spectral pre-processing techniques are required to remove any irrelevant information including noise, uncertainties, variability, interactions and unrecognized features (Lacroix et al., 1997). In this investigation, various pre-processing methods such as first and second derivative, different normalization methods, baseline correction and multiplicative scatter correction (MSC) were tested to find the best one for spectral evaluation. A smoothing of the spectra using a Savitzky-Golay filter with second polynomial order and window size of 15 in combination with standard normal variate (SNV) transformation was applied to reduce the (physical) variability between samples (Karunathilaka et al., 2018) and gave the best results for pre-processing.

For the identification of modified samples, several models were investigated.

### 2.5. Applied models for identification of modified samples

#### 2.5.1. Principal component analysis and regression

A score plot of a principal component analysis (PCA) was employed to investigate if the pattern of reference and modified samples can be separated. Here, the data were analyzed with respect to the variance of the predictor variable by using Unscrambler software. The factor loadings might be able to interpret the dimensions, and the factor scores to

identify the relative positions among the product in a map (Chapman et al., 2001; Karunathilaka et al., 2018). To visualize the differences between sample groups, PCA was carried out involving all the samples. To check this procedure all the wavenumbers of a spectrum were used and PCA score plots were drawn with various combinations of principal components.

With the aim of predicting the difference of reference and modified milk, principal component regression (PCR) was evaluated using Unscrambler software. For this purpose, all of the reference samples were put equal to 0, while the modified ones equal to 1. To select the number of factors utilized for building a PCR model and in order to model the system without overfitting, a cross-validation method was employed. The maximum number of factors used to build the model was selected as 16. RMSE was computed as an indicator of the average error in the analysis for each component which shows how well the model fits the data. The determination coefficient ($R^2$) displaying the fraction of the total variance explained by the models was calculated as well.

#### 2.5.2. Gaussian process regression

For the detection of spectra from modified samples, the Gauss process regression was applied. A very good description of Gauss process regression was given by Rasmussen and Williams (2006). The Gaussian process regression is a supervised machine learning algorithm, which needs spectra of reference as well as modified samples for the calculation of a non-parametric regression model. This probabilistic-based approach uses a Bayesian framework for the regression task. The Matlab function *fitrgp* is used for the model calculation, for prediction the function *predict* has been applied. While the intensity values depending on the wavenumbers were utilized as predictor variables, zero for reference and one for modified samples were considered as response values. To determine the model for the evaluation of new spectra, the covariance function (kernel function) was specified. Three different covariance functions were tested (squared exponential, Matern, and rational quadratic). Pre-processed spectra including all the wavenumbers were used as the inputs of the model. The model was calibrated with the reference and modified milks, while the reference samples were put equal to 0 and modified samples were put equal to 1, the same as PCR. By using the training set the detection limit was specified as the highest prediction value of the reference samples. The prediction set was prepared by modified milks and predicted with the model. Then the determination coefficient ($R^2$) was obtained. RMSE was considered as a criterion to state how well the evaluation techniques can detect modified samples.

#### 2.5.3. Autoencoder

An autoencoder is a special type of feedforward neural network which consists of at least three layers. The input and output layers have the same number of neurons, while there can be just one hidden layer or many with various numbers of neurons. The network learns to map the input pattern to the output pattern through a bottleneck, which is the hidden layer. Therefore, the autoencoder learns an approximation to the identity function. Due to the fact that the number of neurons in the hidden layers are restricted to smaller numbers than the input layer, a compression is carried out. As a learning algorithm, the back-propagation algorithm can be applied. Given sufficient hidden layers, the multilayer feedforward sigmoidal network can approximate virtually any function of interest to any desired degree of accuracy (Sharma et al., 2007). In this investigation, the maximum number of epochs was arbitrarily kept at 10,000. For hidden neurons, the transfer function tanh (x) was applied, which was a rescaled logistic sigmoid function, whereas the identity transfer function was implemented for the output neurons.

In order to detect deviations from reference, the following architecture represented the best results; a three-layer network with 322 input neurons, 10 hidden neurons, and 322 output neurons (just every 10th wavenumber was utilized as the training input and output pattern). The autoencoder was trained with 139 pre-processed spectra taken from various packages of reference milk. During the training step, which lasts roughly 1 min, data was split into the training set and testing set. The training was stopped when the error of the test data increased which can avoid overfitting. To analyze the samples, the sum of squared differences between input and output of the autoencoder was computed by the formula:

$$\mathrm{MSE} = \frac{1}{\boldsymbol{N}\cdot\boldsymbol{n}} \sum_{i=1}^{N} \left(\overrightarrow{\mathrm{x}}_{\mathrm{i,in}} - \overrightarrow{\mathrm{x}}_{\mathrm{i,out}}\right)\left(\overrightarrow{\mathrm{x}}_{\mathrm{i,in}} - \overrightarrow{\mathrm{x}}_{\mathrm{i,out}}\right)^{\mathrm{T}} \quad (1)$$

MSE (considered as autoencoder error) is the mean squared error of applied samples, N is the number of samples, n the number of intensity values in a spectrum, and $\vec{x}_{i,in}$and $\vec{x}_{i,out}$ are the ith spectrum of the input and output of the autoencoder respectively. T denoted the transposed vector. A confidence interval was calculated for the training set using the following equation:

$$\text{Confidence Interval} = \mathrm{MSE} \pm \mathrm{t}(\alpha, \mathrm{f})\cdot \mathrm{s} \quad (2)$$

Here $t()$ is the t distribution, α significance level, $f$ degree of freedom and s is equal to standard deviation of the MSE. 213 pre-processed spectra were employed for the prediction set. The upper boundary of confidence interval of training set with significance level of 5% was utilized as limit to decide if a reference or modified sample is present. The autoencoder decision was "modified sample", when the following prove value p of a new sample was larger than the limit.

$$\mathrm{p} = \left(\overrightarrow{\mathrm{x}}_{\mathrm{in}} - \overrightarrow{\mathrm{x}}_{\mathrm{out}}\right)\left(\overrightarrow{\mathrm{x}}_{\mathrm{in}} - \overrightarrow{\mathrm{x}}_{\mathrm{out}}\right)^{\mathrm{T}} \Big/ n \quad (3)$$

For each sample, the measurement was done several times and after obtaining the prove value for each replication the average $\bar{p}$and standard deviation were calculated. The MATLAB software's proprietary script language (version of 2019b) with Deep Learning Toolbox (version of 13.0) was applied with the aim of developing the autoencoder.

## 3. Results and discussion

### 3.1. Raman spectra

Raman spectrum acquired from the standard milk sample is shown in Fig. 1. The prominent peaks in the Raman spectra are observed at 1650 $cm^{-1}$ (C = C) stretching of RHC = CHR, 1440 $cm^{-1}$ (C–H) bending of -CH2, and 1747 $cm^{-1}$ (C = O) stretching of RC = OOR. The three bands around 1008 $cm^{-1}$ (C–CH3 bending), 1150 $cm^{-1}$ (C–C stretching), and 1525 $cm^{-1}$ (C = C stretching) are attributed to carotenoids. Carotenoids are natural antioxidants and contribute up to 50% of vitamin A activity. In the region of 2800–3100 $cm^{-1}$ of the Raman spectra, bands around 2850 and 2940 $cm^{-1}$ are contributed to the symmetric and asymmetric (C–H) stretching vibrations of the terminal chains of methyl ($CH_2$) and methylene ($CH_3$) groups of aliphatic molecules that exist in edible oils and fats. Also, bands at 3005 $cm^{-1}$ are associated with the symmetric scissoring mode of (= C–H) (El-Abassy et al., 2011). Region of 700–1200 $cm^{-1}$ demonstrates the spectral features of a so-called fingerprint. This area includes bands that are characteristics of C–C skeletal and C–O bond vibrations. The hydrocarbon chains are characterized by a series of bands due to the vibration of skeletal C–C bonds in the wavenumber of 800–900 and 1000–1100 $cm^{-1}$, while C–O bonds have characteristic features in two bands close to 800–970 and 1060–1150 $cm^{-1}$. The carbon chain deformation vibrations are detected in the region of 150–450 $cm^{-1}$ (Baeten et al., 1998).

### 3.2. Detection of changes using principal component analysis and regression

A score plot of principal component analysis and its regression were employed to detect modified samples from the reference. Fig. 2 illustrates a classification among the reference and various modified samples using PCA.

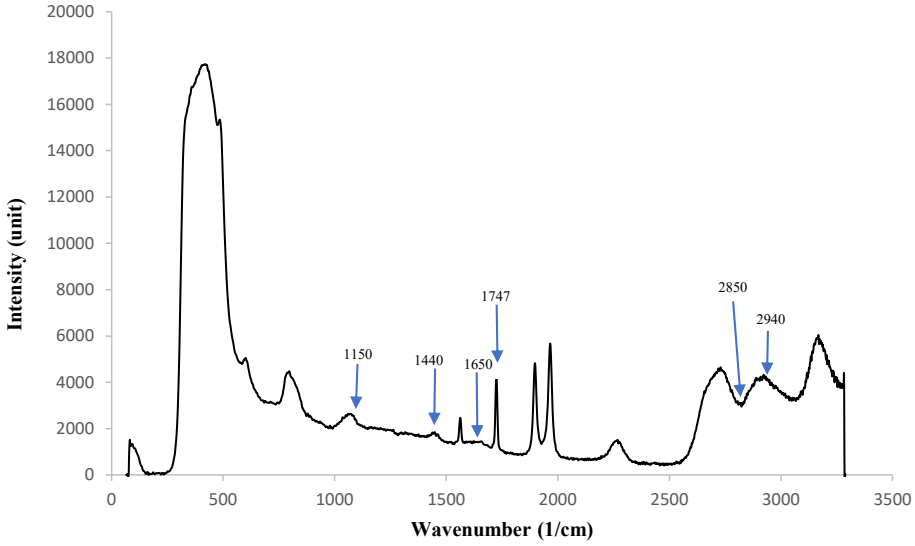


**Fig. 1.** Raw spectrum of reference milk sample acquired by Raman spectrometer using a quartz in-line flow cell.

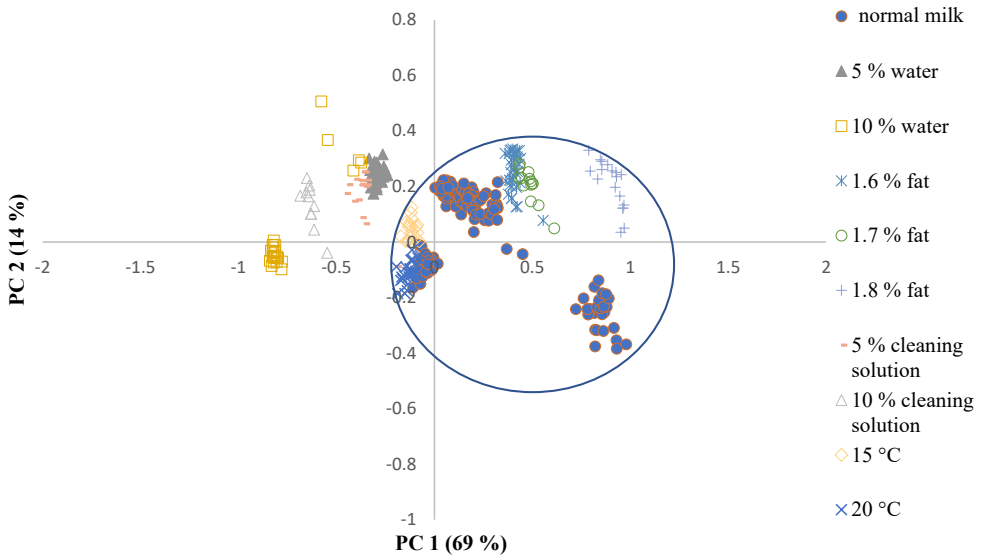


**Fig. 2.** Results of PCA with the entire data set including reference and modified milks using score values for PC 1 and PC 2. Circle indicates the limit of detection.

PCA score plot was drawn with different components, however, PC1 and PC2 separate the samples in some groups a bit better than the other PCs combinations. PC1 covers 69% and PC2 14% of the variance of the pre-processed data set. The region which is defined by the circle presents the reference milk area which can be considered as the limit of detection. Consequently, the samples are situated slightly outside the limit are denoted as the modified samples. As can be seen in Fig. 2, samples with various amount of water and cleaning solution can be distinguished from the reference. The model is not capable to differentiate samples measured at various temperatures. Furthermore, the difference in fat contents cannot be distinguished clearly. It can be concluded that the classification results are not promising. PCA model in combination with Raman spectroscopy has been recently used in order to discriminate infant gender on the basis of milk fat content and shown much better results. In total 50 milk samples were used in this study, out of which 20 have male and 30 have female infant babies. It was observed that PCA was able to clearly separate the two sets of classes on the basis of features obtained from the Raman spectra (Ullah et al., 2017).

With the aim of predicting the reference and modified samples, PCR was evaluated. Fig. 3 a presents results obtained by principal component regression for calibration of the model. For the PCR model, a calibration set with 16 principal components was used, including reference samples (whose y values were set to 0) and modified milks (whose y values were set to 1).

Fig. 3b presents the principal component regression model for predicting modified samples and reference samples.

As can be seen in Fig. 3a showing the calibration set after implementing PCR, reference samples spread out from −0.5 to 0.9 and modified samples from 0.2 to 1.5. Subsequently, a threshold for group classification cannot be selected. In addition, Fig. 3b plotting predicted values against reference values demonstrates the $R^2$ equal to 0.7 and RMSE of 0.6 which implies the inability of PCR to predict reference and modifies samples correctly.

He et al. (2019) mentioned that the baseline correction algorithms can be employed for suppressing fluorescence background, while the Savitzky-Golay filter is commonly used to improve signal-to-noise. Moreover, normalization algorithms (standard normal variate, multiplicative scatter correction, peak normalization) are applied for evaluating real samples. More approaches and their applicability in Raman analysis have been stated elsewhere (Lohumi et al., 2017; De Luca et al.,

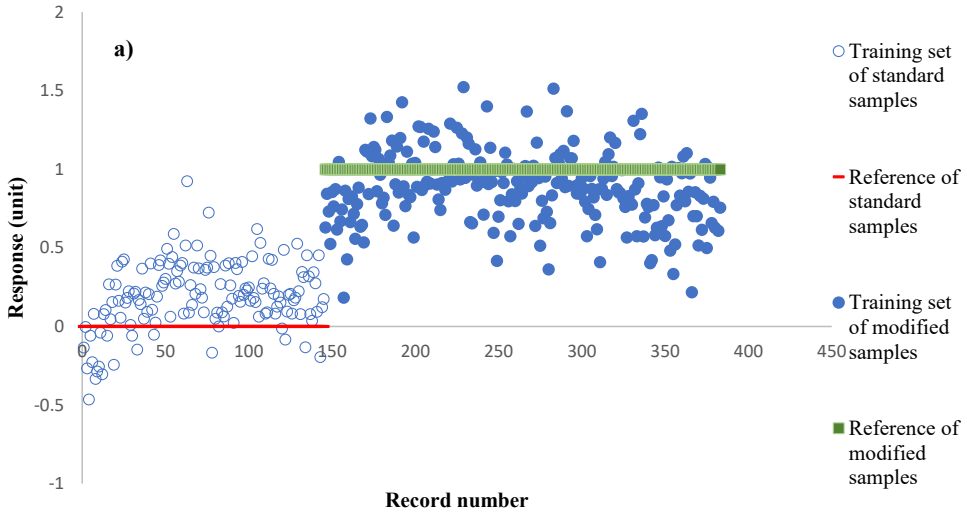


**Fig. 3a.** Calibration data according to the principle component regression.

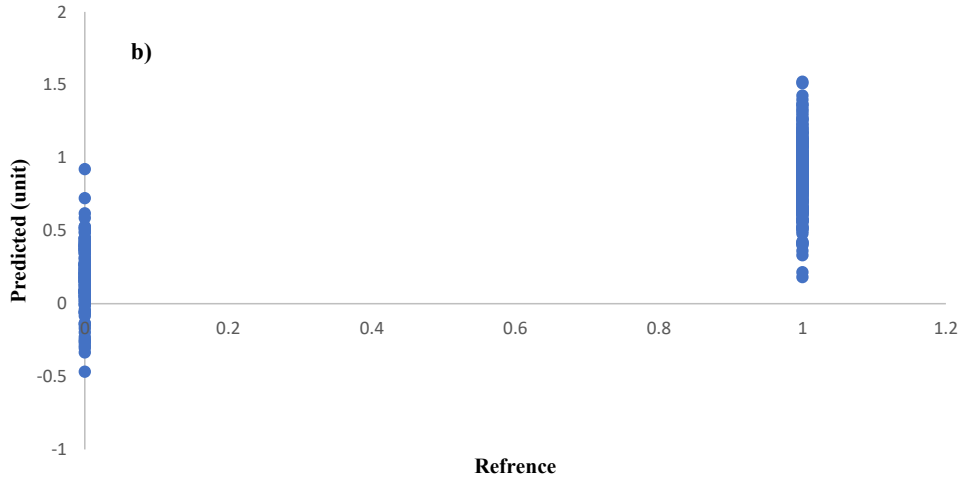


**Fig. 3b.** Results of prediction by using principle component regression and cross validation.

2015).

### 3.3. Detection of changes by using Gaussian process regression

In pursuance of finding better results, a Gaussian process regression model using cross-validation was employed to determine changes. Fig. 4 represents the results of the calibration set and prediction set acquired by the Gaussian process regression model using pre-processed data. The determination coefficient $R^2$ and RMSE acquired by this model are 0.97 and 0.09 respectively. The limit of detection obtained by this procedure is equal to 0.6.

Fig. 5 displays the results of the prediction set of Gaussian process regression including the percentage of values higher and lower than the limit of detection.

The averages and standard deviations of predicted values using GPR are illustrated in Table 1.

In this step, reference samples were put equal to zero and modified ones equal to one and as can be observed in Fig. 4 a the "Reference'' line is implying this procedure. According to Fig. 4 a, in the calibration set responses of the reference samples are spreading from −0.09 to 0.5 and modified samples from 0.7 to 1.1. The limit of detection can be defined by the value of 0.6 which can separate the modified and reference milks quite good. It implies the fact that each predicted value higher than the limit is considered as a modified sample. In regard to Fig. 4 b, the sample with 5% water is mostly higher than the limit. By a closer look at Figs. 5 and 94.9% of replications of this sample are higher than the value of 0.6 considered as the limit of detection. Subsequently, regarding the average of 0.7 for this sample, 5% water is detectable by this method. Although, all the replications of milks with 10% water are higher than the limit. As can be seen in Fig. 4 b, all the predicted values for milks with various fat contents are higher than the limit, which implies the ability of the Gaussian process regression model to distinguish the difference of 0.1% fat in the milk. In addition, this model is also able to detect samples with 5% and 10% cleaning solution because of the higher values than the detection limit. Milks were measured at 15 °C are spreading around the limit. In the mentioned sample 73.7% of repetition are higher than 0.6 showing an average of 0.7, which implies the capability of GPR for detection. Despite samples measured at 15 °C, all the sample measured at 20 °C are obviously above the limit of detection. The model predicts the reference samples with values lower than the limit.

Generally, by implementing the combination of Raman spectrometer and Gaussian process regression model all the modified sample can be clearly distinguished from the reference samples. However, pre-processing by Savitzky-Golay filter and SNV improved the results.

Fu et al. (2019) applied various methods to predict the diet nutrients

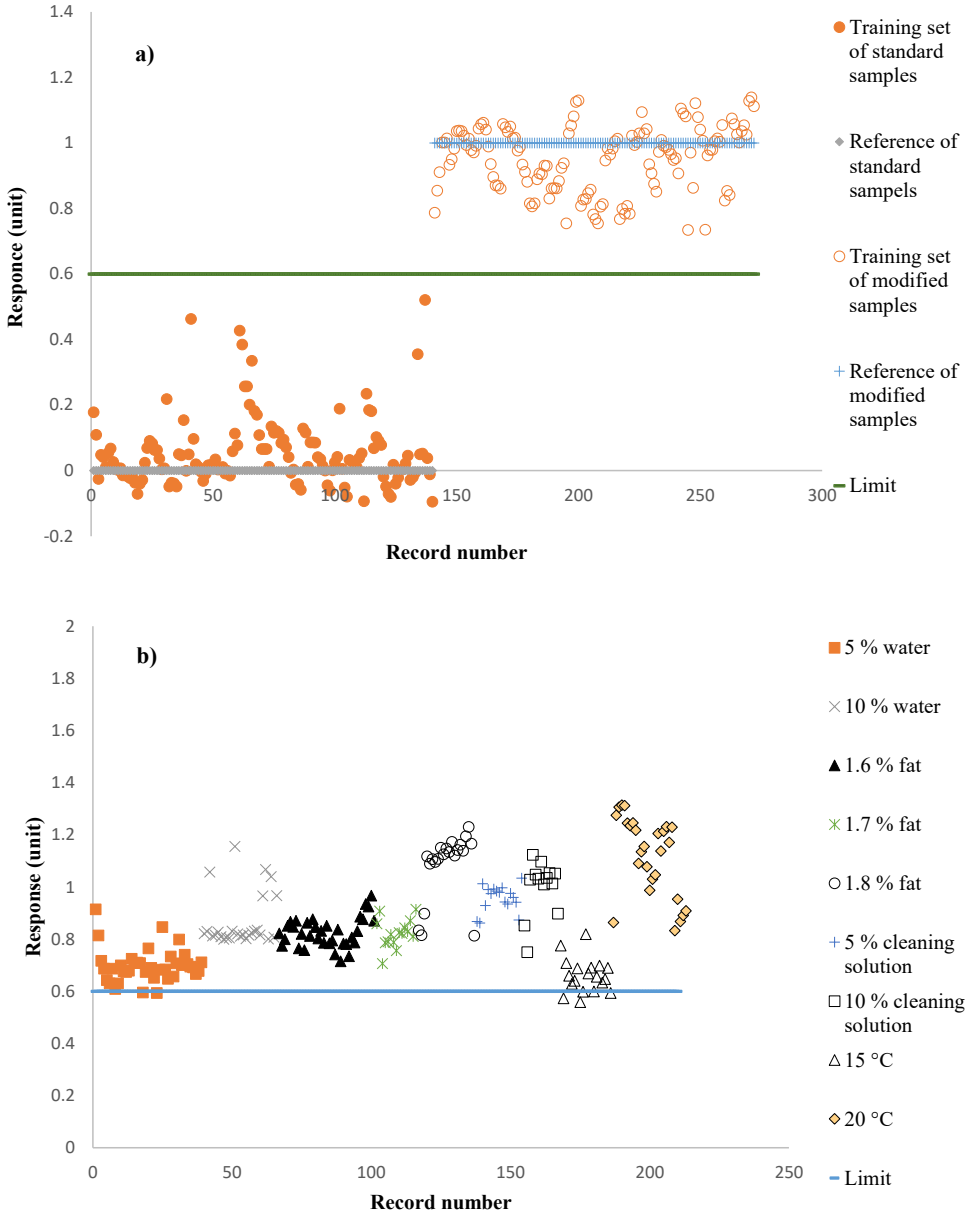


**Fig. 4.** Graph for calibration set (4.a) and prediction set (4.b) acquired by Gaussian process regression model using pre-processed spectra.

digestibility of dairy cows and the prediction results specified that the GPR method is superior to other conventional techniques. A large number of researches have represented that the GPR method generally achieves better modelling and prediction performance than other statistical prediction techniques (Kong et al., 2018).

### 3.4. Detection of changes by autoencoder

To investigate the effectiveness of the autoencoder on the detection of deviations from standard during the milk processing, pre-processed data was employed as the inputs. In addition, for the training set, the average and standard deviation of autoencoder errors were calculated, 2.7E-04 and 7.0E-05 respectively. The limit of detection, which was computed by using the upper boundary of the confidential interval with the significance level of 5%, is equal to 4.1E-04. Consequently, samples that show higher values than the limit of detection is considered as the modified samples.

Fig. 6 represents the autoencoder errors of the training set and prediction set by using pre-processed data.

The averages and standard deviations of autoencoder errors in the prediction set are displayed in Table 1. Fig. 7 represents the percentage of autoencoder errors that are lower and higher than 4.1E-04, which is the limit of detection.

In regard to Fig. 7, the autoencoder can distinguish 5% water in the milk as the percentage of values higher than the limit is 81.1%. According to Table 1, various replications of the mentioned sample show the average of 5.3E-04 ± 1.8E-04 which is higher than the limit value implies the ability of the autoencoder for detection. Confirming to Fig. 6 b, all the repetitions of milk with 10% water are higher than the limit which states that it can be distinguished perfectly by this method. As can be seen in Fig. 6 b, samples with 1.6% fat spread out around the limit. By closer look at Fig. 7, it can be defined that 42.9% of predicted values of 1.6% fat milk are higher than the limit while it is 66.7% for samples with 1.7% fat. In contrast, all the replications for milks with 1.8% fat show values higher than the limit. Overall, while the autoencoder errors of milk with 1.6% fat are spreading around the limit, the average of 5.5E-04 is still higher than the limit. Consequently, the difference of 0.1% fat is detectable by using the autoencoder. According to Fig. 6 b, various

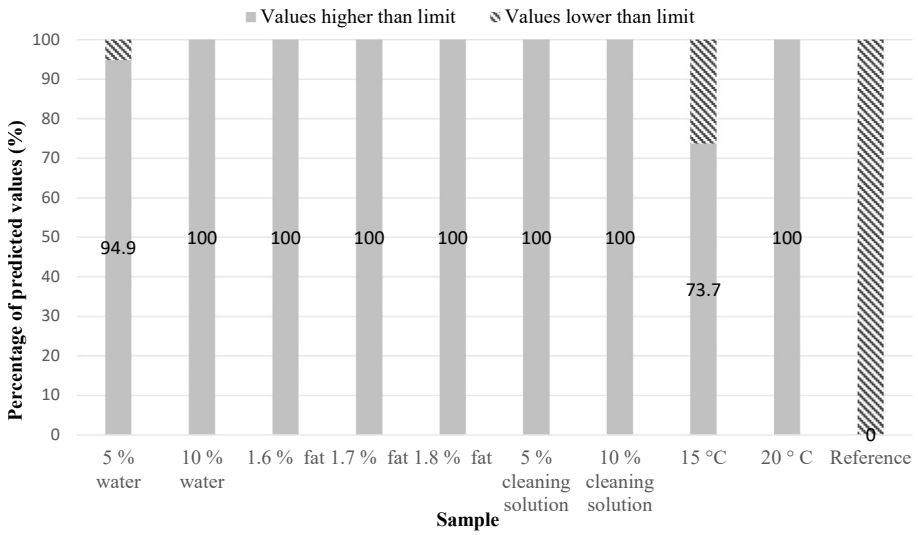


**Fig. 5.** Percentage of predicted values higher and lower than the limit of detection using Gaussian process regression.

**Table 1**
Results of prediction set acquired by the autoencoder and Gaussian process regression model. Limit of detection for GPR is 0.6 and for autoencoder is equal to 4.1E-04.

| Samples | Average of autoencoder errors (unit) | Standard deviation of autoencoder errors (unit) | Average of predicted values using GPR (unit) | Standard deviation of predicted values using GPR (unit) |
|---|---|---|---|---|
| 5% water | 5.3E-04 | 1.8E-04 | 0.7 | 0.06 |
| 10% water | 8.0E-04 | 2.0E-04 | 0.9 | 0.10 |
| 1.6% fat | 5.5E-04 | 7.8E-04 | 0.8 | 0.06 |
| 1.7% fat | 1.7E-03 | 4.5E-03 | 0.8 | 0.05 |
| 1.8% fat | 3.2E-03 | 7.7E-03 | 1.1 | 0.12 |
| 5% cleaning solution | 9.0E-04 | 1.3E-03 | 1.0 | 0.05 |
| 10% cleaning solution | 3.0E-03 | 5.8E-03 | 1.0 | 0.10 |
| 15 °C | 4.2E-04 | 7.9E-05 | 0.7 | 0.07 |
| 20 °C | 1.0E-03 | 2.8E-03 | 1.1 | 0.15 |
| Reference | 3.39E-04 | 3.06E-05 | 0.02 | 0.12 |

concentrations of cleaning solution in the milk (5% and 10%) are obviously higher than the limit. Therefore, it signifies the capability of the autoencoder to identify 5% cleaning solution in the milk. The results of reference milk measured at 15 °C demonstrate that 42.1% of autoencoder errors are higher than the limit which shows the average of 4.2E-04 in Table 1. Despite the average value is very close to the limit of detection, it is still higher implying the capability of the autoencoder for considering the sample as a modified one. For samples, which were measured at 20 °C the autoencoder is functional to detect changes as 79.2% of replications with an average of 1.0E-03 are higher than the limit of detection.

Totally, the combination of the autoencoder with the Raman spectrometer shows promising results for the identification of modified samples. The pre-processed spectra by the smoothing prior to calculating the SNV transformation can affect the mentioned network to distinguish deviations from standard production perfectly. Consequently, it is capable to distinguish the 5% water and cleaning solution in the milk while Vasafi et al. (2021) illustrated that using the autoencoder in combination with a NIR spectrometer is not able to detect 5% cleaning solution (similar cleaning solution) in the milk processing. Difference of 0.1% in fat content and 5 °C difference in the measurement temperature are also detectable by utilizing the autoencoder.

To compare the results of the autoencoder and Gaussian process regression model, it can be mentioned that both are showing promising results and are able to detect various changes. By comparing Fig. 4.b and 6 b, it can be seen that GPR can detect more modified signals than the autoencoder. However, the model must be calibrated with all the possible changes before the usage which can be considered as the disadvantage of this technique. Consequently, the autoencoder learns how a regular signal looks like and upon this, changes can be detected. After a training phase, it is ready to use. It will always be attentive and never become tired and is highly recommended.

## 4. Conclusion

Monitoring of milk processing using a process analyser helps the companies to ensure the quality and safety of products before filling and packaging. Otherwise, samples have to be taken from the packages and analyzed. Establishing a novel procedure, which can detect deviations from standard processing (normal data) in a by-pass, helps to minimize the risks of recalls of products with quality defects. In this contribution, we used the contact-free measurement *via* Raman spectrometer in a by-pass combined with data analysis of the spectra to identify deviations of modified milk from reference milk. The principal component analysis and regression of the spectral data did not show a promising prediction as the related $R^2$ was 0.7. Gaussian process regression with $R^2$ of 0.97 was capable to predict the changes almost perfectly. In addition, results represented the proper functionality of the autoencoder for detecting deviations from standard milk production. Both of these innovative procedures displayed promising results including the detection of 5% water and cleaning solution, the difference of 0.1% in fat content, and variation of 5 °C in the measurement temperature. In summary, to compare the results of these two techniques, Gaussian process regression can detect more modified signals. However, the autoencoder can be trained with just standard milk samples, no preparation of modified samples are necessary, and immediately the supervision of the process can be started. Therefore, an autoencoder is suggested for such applications.

## CRediT authorship contribution statement

**Pegah Sadeghi Vasafi:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Writing – original draft. **Jörg Hinrichs:** Supervision, Writing – review & editing. **Bernd**

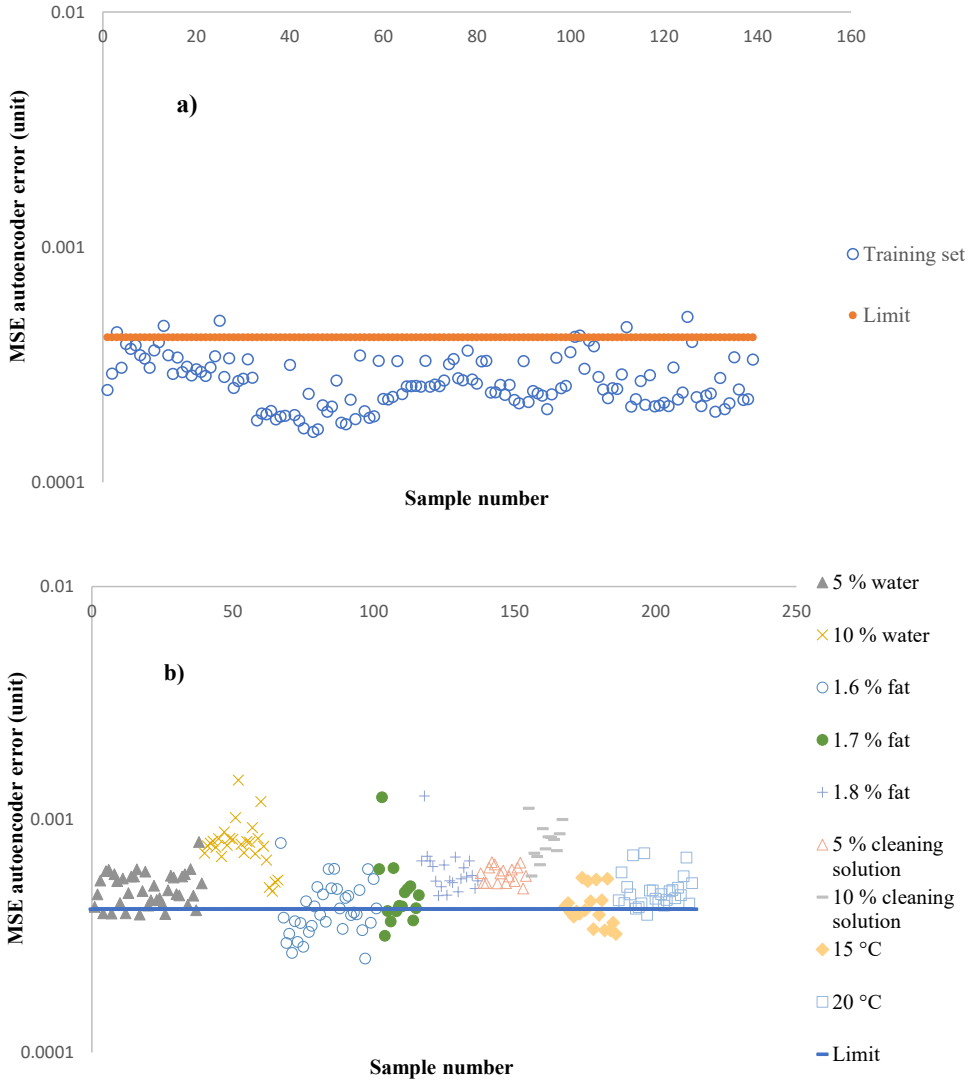


**Fig. 6.** Graph for training set showing the mean squared error (MSE) of input and output of pre-processed spectrum (6.a) and graph for prediction set showing MSE of input and output by using autoencoder (6.b). Limit is equal to upper boundary of confidential interval of training set with significance level of 5% which is shown by straight line.

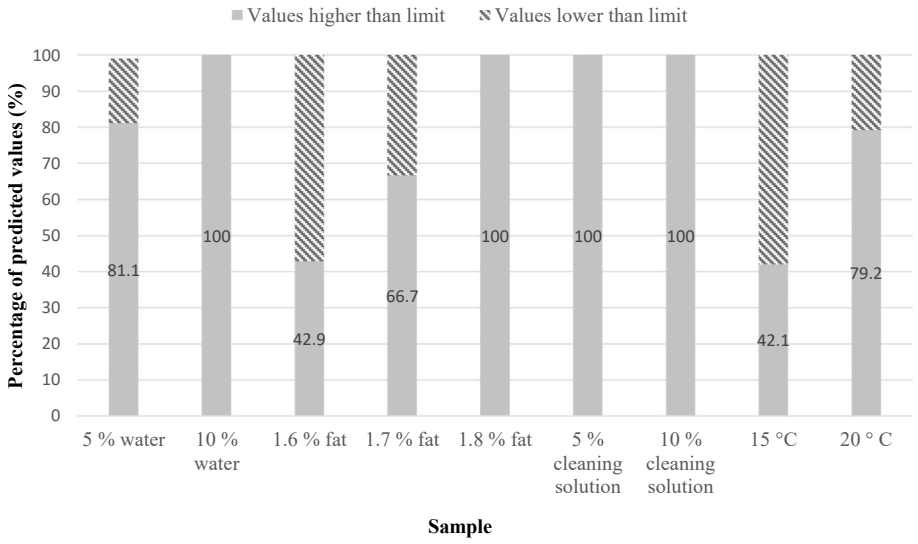


**Fig. 7.** Percentage of predicted values higher and lower than the limit of detection using the autoencoder.

**Hitzmann:** Conceptualization, Supervision, Writing – review & editing, Project administration.

## Declaration of competing interest

Please check the following as appropriate:

All authors have participated in (a) conception and design, or analysis and interpretation of the data; (b) drafting the article or revising it critically for important intellectual content; and (c) approval of the final version.

This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue.

The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript.

The following authors have affiliations with organizations with direct or indirect financial interest in the subject matter discussed in the manuscript:

## Acknowledgments

This IGF project of the FEI is/was supported via AiF (20200N) within the programme for promoting the Industrial Collective Research (IGF) of the German Ministry of Economics and Energy (BMWi), bases on a resolution of the German Parliament for which the authors are grateful.

## References

Acar-Soykut, E., Tayyarcan, E. K., & Boyaci, I. H. (2018). A simple and fast method for discrimination of phage and antibiotic contaminants in raw milk by using Raman spectroscopy. *Journal of Food Science and Technology, 55*(1), 82–89.

Baeten, V., Hourant, P., Morales, M. T., & Aparicio, R. (1998). Oil and fat classification by FT-Raman spectroscopy. *Journal of Agricultural and Food Chemistry, 46*(7), 2638–2646. https://doi.org/10.1021/jf9707851

Bernuy, B., Meurens, M., Mignolet, E., & Larondelle, Y. (2008). Performance comparison of UV and FT-Raman spectroscopy in the determination of conjugated linoleic acids in cow milk fat. *Journal of Agricultural and Food Chemistry, 56*(4), 1159–1163. https://doi.org/10.1021/jf703712w

Blanpain-Avet, P., Hedoux, A., Guinet, Y., Paccou, L., Petit, J., Six, T., & Delaplace, G. (2012). Analysis by Raman spectroscopy of the conformational structure of whey proteins constituting fouling deposits during the processing in a heat exchanger. *Journal of Food Engineering, 110*(1), 86–94. https://doi.org/10.1016/j.jfoodeng.2011.12.005

Chapman, K. W., Lawless, H. T., & Boor, K. J. (2001). Quantitative descriptive analysis and principal component analysis for sensory characterization of ultra pasteurized milk. *Journal of Dairy Science, 84*(1), 12–20. https://doi.org/10.3168/jds.S0022-0302(01)74446-3

De Luca, A. C., Dholakia, K., & Mazilu, M. (2015). Modulated Raman spectroscopy for enhanced cancer diagnosis at the cellular level. *Sensors, 15*(6), 13680–13704. https://doi.org/10.3390/s150613680

Dijkstra, R. J., Ariese, F., Gooijer, C., & Brinkman, U. T. (2005). Raman spectroscopy as a detection method for liquid-separation techniques. *TrAC Trends in Analytical Chemistry, 24*(4), 304–323. https://doi.org/10.1016/j.trac.2004.11.022

El-Abassy, R. M., Eravuchira, P. J., Donfack, P., Von der Kammer, B., & Materny, A. (2011). Fast determination of milk fat content using Raman spectroscopy. *Vibrational Spectroscopy, 56*(1), 3–8. https://doi.org/10.1016/j.vibspec.2010.07.001

Fu, Q., Shen, W., Wei, X., Zheng, P., Xin, H., & Zhao, C. (2019). Prediction of the diet nutrients digestibility of dairy cows using Gaussian process regression. *Information Processing in Agriculture, 6*(3), 396–406. https://doi.org/10.1016/j.inpa.2018.11.005

Gong, D., Liu, L., Le, V., Saha, B., Mansour, M. R., Venkatesh, S., & Van Den Hengel, A. (2019). *Memorizing normality to detect anomaly: Memory-augmented Deep autoencoder for unsupervised anomaly detection.* Presented at the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 1705–1714.

He, H., Sun, D. W., Pu, H., Chen, L., & Lin, L. (2019). Applications of Raman spectroscopic techniques for quality and safety evaluation of milk: A review of recent developments. *Critical Reviews in Food Science and Nutrition, 59*(5), 770–793. https://doi.org/10.1080/10408398.2018.1528436

Karunathilaka, S. R., Yakes, B. J., He, K., Brückner, L., & Mossoba, M. M. (2018). First use of handheld Raman spectroscopic devices and on-board chemometric analysis for the detection of milk powder adulteration. *Food Control, 92*, 137–146. https://doi.org/10.1016/j.foodcont.2018.04.046

Kong, D., Chen, Y., & Li, N. (2018). Gaussian process regression for tool wear prediction. *Mechanical Systems and Signal Processing, 104*, 556–574. https://doi.org/10.1016/j.ymssp.2017.11.021

Lacroix, R., Salehi, F., Yang, X. Z., & Wade, K. M. (1997). Effects of data pre-processing on the performance of artificial neural network for dairy yield prediction and cow culling classification. *Transactions of the ASAE, 40*, 839–846. American Society of Agricultural Engineers.

Li, M., Chen, J., Xu, J., Fu, S., & Gong, H. (2015). Determination of lactose in milk by Raman spectroscopy. *Analytical Letters, 48*(8), 1333–1340. https://doi.org/10.1080/00032719.2014.979358

Liu, Z., Huang, M., Zhu, Q., Qin, J., & Kim, M. S. (2019). Packaged food detection method based on the generalized Gaussian model for line-scan Raman scattering images. *Journal of Food Engineering, 258*, 9–17. https://doi.org/10.1016/j.jfoodeng.2019.04.005

Lohumi, S., Kim, M. S., Qin, J., & Cho, B. K. (2017). Raman imaging from microscopy to macroscopy: Quality and safety control of biological materials. *TrAC Trends in Analytical Chemistry, 93*, 183–198. https://doi.org/10.1016/j.trac.2017.06.002

Moros, J., Garrigues, S., & de la Guardia, M. (2007). Evaluation of nutritional parameters in infant formulas and powdered milk by Raman spectroscopy. *Analytica Chimica Acta, 593*(1), 30–38. https://doi.org/10.1016/j.aca.2007.04.036

Rajapandiyan, P., Tang, W. L., & Yang, J. (2015). Rapid detection of melamine in milk liquid and powder by surface-enhanced Raman scattering substrate array. *Food Control, 56*, 155–160. https://doi.org/10.1016/j.foodcont.2015.03.028

Raman, C. V., & Krishnan, K. S. (1928). A new type of secondary radiation. *Nature, 121*(3048), 501–502.

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning. MIT Press*, Massachusetts Institute of Technology.

Sharma, A. K., Sharma, R. K., & Kasana, H. S. (2007). Prediction of first lactation 305-day milk yield in Karan Fries dairy cattle using ANN modeling. *Applied Soft Computing, 7*(3), 1112–1120. https://doi.org/10.1016/j.asoc.2006.07.002

Stefanov, I., Baeten, V., Abbas, O., Colman, E., Vlaeminck, B., De Baets, B., & Fievez, V. (2011). Determining milk isolated and conjugated trans-unsaturated fatty acids using Fourier transform Raman spectroscopy. *Journal of Agricultural and Food Chemistry, 59*(24), 12771–12783.

Ullah, R., Khan, S., Khan, A., Saleem, M., Ali, H., Bilal, M., & Ahmed, M. (2017). Infant gender based differentiation in concentration of milk fats using near infrared Raman spectroscopy. *Journal of Raman Spectroscopy, 48*(3), 363–367. https://doi.org/10.1002/jrs.5047

Vasafi, P. S., Paquet-Durand, O., Brettschneider, K., Hinrichs, J., & Hitzmann, B. (2021). Anomaly detection during milk processing by autoencoder neural network based on near-infrared spectroscopy. *Journal of Food Engineering, 299*, 110510. https://doi.org/10.1016/j.jfoodeng.2021.110510

Vaskova, H., & Buckova, M. (2016). Measuring the lactose content in milk. *MATEC Web of Conferences, 76*, 4.

Wang, Y., & Chaib-draa, B. (2016). KNN-based Kalman filter: An efficient and non-stationary method for Gaussian process regression. *Knowledge-Based Systems, 114*, 148–155. https://doi.org/10.1016/j.knosys.2016.10.002

Wang, M. C., Han, X., Sun, Y. M., Sun, Q. Y., & Chen, X. G. (2021). Study on soft sensor modeling method for sign of contaminated fermentation broth in Chlortetracycline fermentation process. *Preparative Biochemistry & Biotechnology, 51*(1), 76–85. https://doi.org/10.1080/10826068.2020.1793173

## 3.3. Comparison of various classification techniques for supervision of milk processing

By Pegah Sadeghi Vasafi, Bernd Hitzmann. Published in Engineering in Life Sciences. October 2021. In press.

Received: 5 August 2021 | Revised: 18 October 2021 | Accepted: 18 October 2021

DOI: 10.1002/elsc.202100098

Engineering in Life Sciences

RESEARCH ARTICLE

# Comparison of various classification techniques for supervision of milk processing

**Pegah Sadeghi Vasafi** | **Bernd Hitzmann**

Process Analytics and Cereal Science, University of Hohenheim, Stuttgart, Germany

**Correspondence**
Pegah Sadeghi Vasafi, Process Analytics and Cereal Science, University of Hohenheim, Garbenstr. 23, 70599, Stuttgart, Germany.
Email: Pegah.sadeghivasafi@uni-hohenheim.de

**Abstract**

Detecting the types of anomalies that can occur throughout the milk processing process is an important task since it can assist providers in maintaining control over the process. The Raman spectrometer was used in conjunction with several classification approaches—linear discriminant analysis, decision tree, support vector machine, and k nearest neighbor—to establish a viable method for detecting different types of anomalies that may occur during the process—temperature and fat variation and added water or cleaning solution. Milk with 5% fat measured at 10°C was used as the reference milk for this study. Added water, cleaning solution, milk with various fat contents and different temperatures were used to detect abnormal conditions. While decision trees and linear discriminant analysis were unable to accurately categorize the various type of anomalies, the k nearest neighbor and support vector machine provided promising results. The accuracy of the support vector machine test set and the k nearest neighbor test set were 81.4% and 84.8%, respectively. As a result, it is reasonable to conclude that both algorithms are capable of appropriately classifying the various groups of samples. It can assist milk industries in determining what is wrong during milk processing.

**KEYWORDS**
anomaly detection, classification methods, milk processing, Raman spectroscopy

## 1 | INTRODUCTION

Milk and dairy products are well-known for their benefits to human health. Milk is one of the main components of the human diet and a universal source of nutrients for protein, lactose, vitamins, minerals, and fats [1]. To minimize the production problems, large resource investments are required [2]. Using fast spectroscopy to detect product defects online is beneficial to dairy producers as it can readjust product characteristics or redirect product flow during the production process. During the processing of milk, some abnormal changes—fat and temperature variation, added water and cleaning solution—can happen which threaten the quality and safety of final products. As the fat content of milk is usually set and they are classified by their amount of fat, fat concentration should be consistent and correct during production. Also, the pilot plant is cleansed with cleaning solution and water after production and if some part of the chemical stuff remains in the

**ABBREVIATIONS:** kNN, k nearest neighbor; SVM, support vector machine; UHT, ultra-high temperature processing milk

production line, it creates numerous safety and quality concerns. Moreover, the temperature is the most important option which has to be controlled, as milk is a heat-treated product. As a result, controlling these changes is a beneficial task [2, 3, 4]. Therefore, controlling the process online is a vital sector that helps a company avoid suffering. In this case, not only detecting the abnormal changes would be an advantage but also it is very important to understand what exactly happened in the processing steps. Therefore, a predictive tool based on online measurement data is needed to monitor every stage of production [3].

Raman spectroscopy has great potential in such applications due to its quick and easy measurement. It has great potential in food quantification and has been applied to food science, especially for dairy technology. A fast-screening approach for detecting melamine in milk powder with laser Raman spectrometry was developed by Cheng et al. [5] which a detection limit of 0.13% and a good partial least squares (PLS) analysis model were obtained. McGoverin et al. [6] represented the efficiency of Raman spectroscopy at quantifying the protein and fat within skim and whole milk powders. Also, it was effective in the identification of additives such as calcium carbonate. Taking barista foam as an example, the applicability of Raman spectroscopy as a product application parameter index was studied. In order to evaluate the applicability of a purely online system, principal component analysis was used to evaluate the advantages of Raman spectroscopy [07].

Machine learning includes the use of mathematics, statistics, and calculation methods, with the goal of finding effective and accurate classification algorithms. Machine learning algorithms for classification have been successfully used in many different applications, such as food science. The classification problem learning step usually starts with a set of labeled examples containing a training set and a test set. Ciosek et al. [8] classified milk with the use of support vector machine networks. The numerical results of the recognition of milk made differently and with variable fat content have proven to be quite good. A research was done to investigate the use of the least-squares support vector machine (LS-SVM) as an alternative multivariate calibration method for the simultaneous quantification of some common adulterants (starch, whey or sucrose) found in powdered milk samples, using near-infrared spectroscopy with direct measurements by diffuse reflectance and showed promising results [9]. For the automated microbiological quality evaluation of pasteurized vanilla cream, the performance of Fourier transform infrared (FTIR) spectroscopy with support vector machine analysis, was evaluated by Lianou et al. [10]. In the other study, Raman spectral data of milk samples of different species were used for multi-class classification using a dimensionality reduction technique in combination with a random forest (RF) classifier. With an average accuracy of about 93.7%, precision of 94%, specificity of 97%, and sensitivity of 93%, the suggested technique indicated a considerable potential for a distinction between milk samples of different species [11]. De Lima et al. [12] presented a rapid method for discrimination between lactose and lactose-free UHT milks using NIRS combined with multivariate classification methods. Among the classification models developed, LDA (linear discriminant analysis) models were more parsimonious due to the use of fewer variables. Although, k nearest neighbor (kNN) classification model was developed to classify control from adulterated milk samples and adulterated milk samples based on the level of adulteration. The results illustrated quite satisfactory predictability, with sensitivity and specificity ranging from 0.66 to 1 [13]. A decision tree (DT) model was utilized to detect post-calving diseases based on rumination, activity, milk yield, BW and voluntary visits to the milking robot. The overall accuracy of the model was 78%, with a specificity of 87% and a sensitivity of 69%, suggesting its practical value [14]. Using Raman spectroscopy Vasafi et al. [15] could demonstrate that Gaussian process regression as well as autoencoder were able to distinguish between reference milk and manipulated milk, but could not identify which manipulation took place.

The main goal of this research was to develop a suitable technique based on data obtained by Raman spectrometer, not only for detecting various changes that can happen during the processing of milk—changes in fat, temperature, added water or contamination of cleaning solution—but also to label which change happened. In this contribution, various classification methods—linear discriminant analysis, decision tree, support vector machine, k nearest neighbor—were tested to find the best method of detection.

**PRACTICAL APPLICATION**

Detection of anomalies during the processing of food would be helpful; however, it is even more important to determine the type of anomalies that happened during the process. It can help companies not only to understand the existence of anomalies in the process but also help them to find the type of it. As a result, classification can assist industries to detect the type of anomalies that happened during milk processing as fast as possible and avoid being suffered.

**TABLE 1** Preparation of the abnormal samples

| Sample | Preparation method |
|---|---|
| 1.6% fat | Mixing 5 mL of 3.5% fat milk and 95 mL of 1.5% fat milk |
| 1.7% fat | Mixing 10 mL of 3.5% fat milk and 90 mL of 1.5% fat milk |
| 1.8% fat | Mixing 15 mL of 3.5% fat milk and 85 mL of 1.5% fat milk |
| 5% cleaning solution | Adding 5 mL of common cleaning solution to 95 mL of reference milk |
| 5% water | Adding 5 mL of water to 95 mL of reference milk |
| 10% cleaning solution | Adding 10 mL of common cleaning solution to 90 mL of reference milk |
| 10% water | Adding 10 mL of water to 90 mL of reference milk |
| 15°C | Reference samples were measured after heating up to 15°C |
| 20°C | Reference samples were measured after heating up to 20°C |

## 2 | MATERIALS AND METHODS

### 2.1 | Sample preparation

In this contribution, ultra-high temperature processing milk (UHT) with two different fat content of 1.5% and 3.5% were utilized from the brand of "Weihenstephan", Germany. Samples were kept at the temperature of 5°C before opening the packages. The 1.5% fat milk was utilized as a reference sample. A sample with a concentration of 1.6% fat was prepared by mixing 5 mL of 3.5% fat milk and 95 mL of 1.5% fat milk. In addition, kinds of milk with 1.7% and 1.8% fat content were created by the same procedure. One milliliter of a cleaning solution named "Anti-Germ clean A-N 30″ was diluted with 99 of mL water in order to prepare a common cleaning solution. Therefore, different concentrations of water and cleaning solution (0.05 and 0.1 L/L) were added to 1.5% fat milk. All the samples were measured at 10°C. Finally, 1.5% fat milk was measured after heating up to 15°C and 20°C. The purpose behind this work was to find a proper procedure that can clarify what kind of changes happened during the milk processing. Table 1 presents how various modified samples were created.

### 2.2 | Raman spectroscopy

In this study, an Inno-Spec Raman 785 spectrometer (Inno-Spec GmbH, Germany) with a laser excitation wavelength of 785 nm was used to measure the samples. All Raman spectra included scanning with a resolution of 1 $cm^{-1}$ in the spectral range of 65-3290 $cm^{-1}$. The integration time (IT) used was 20 s. A high-quality quartz flow cell with a channel length of 1 mm was used for the measurement. To keep the measurement temperature stable at 10°C, the quartz flow cell was connected to the milk source in a cold-water bath. The flow cell had a capacity of less than 1 mL and was placed 12 mm away from the laser. For each sample, 70% of data were used as the training set while 30% were employed for the test set. One hundred fifty spectra were used for the reference sample and for each modified sample on average 20 spectra were used.

### 2.3 | Pre-processing

Preprocessing has been deemed essential for subsequent data mining tasks and has been determined to be an indispensable part of spectral data analysis. In fact, it has been shown that classification and quantitative models developed based on pre-processed data generally perform better than models based on raw data. Pre-processing includes outlier rejection, normalization, filtering, detrending, conversion, folding, and feature selection. The purposes behind spectral preprocessing are better spectral interpretability, greater robustness, and higher precision of post-classification or quantitative analysis [16]. Therefore, to improve the results, the following preprocessing steps were completed and tested: baseline correction, normalization, multiplicative scatter correction and derivative. Finally, before calculating the standard normal variable (SNV), a Savitzky Golay filter with a second-order polynomial and a window size of 15 was used to smooth the spectrum. SNV belongs to a group of scatter correction preprocessing methods and can reduce physical variability between samples [17].

### 2.4 | Classification algorithms

#### 2.4.1 | Overview

Given an unlabeled sample, the classification problem involves determining which class it belongs to, based on a training data set with known class variables. For showing the results of classification a confusion matrix was used. A confusion matrix is an n×n table that summarizes how successful a classification model's predictions were; that is, the correlation between the class and the model's classification. One axis of a confusion matrix is the class that the model predicted, and the other axis is the actual class. n represents the number of different classes. Four concepts were introduced called true positives, true negatives, false

positives, and false negatives. A true positive is an outcome where the model correctly predicts the reference samples class. Similarly, a true negative is an outcome where the model correctly predicts the modified samples class. In addition, a false positive is an outcome where the model incorrectly predicts the reference class and a false negative is an outcome where the model incorrectly predicts the modified class. In order to evaluate the classification model, its accuracy was calculated. Classification accuracy is the ratio of correct predictions to total predictions made. In the case of unbalanced data sets, the precision of the classification alone is not the best indicator to evaluate the classifier. Various other performance indicators can be used to gain a more complete understanding of the function of the classifier. The confusion matrix contains enough information to calculate various performance indicators—precision, specificity, and recall. Recall or sensitivity is the metric that measures the accuracy on the positive instances, it can be calculated as true positive/(true positive + false negative). Specificity measures the accuracy on the negative instances and can be defined as true negative/(true negative + false positive). Precision is another metric which is the ratio of true positives to the total of the true positives and false positives [18].

### 2.4.2 | Linear discriminant analysis

LDA was first proposed by Fisher in [19]; today, it is still a complete statistical-based pattern classification method. Discriminant function analysis is a dimensionality reduction technique commonly used for supervised classification problems. It is used to model differences in groups, such as separating two or more classes. It can be used to project features in space from high-dimensional to low-dimensional ones. Therefore, it focuses on the separation ability among the classes. In this technique, a new axis is created based on maximizing the distance between the means of each category and minimizing the variation within each category [20]. For the training set, each class was named by a number and the test set was predicted by using the function of predict in MATLAB with the full covariance function.

### 2.4.3 | Decision tree

By its simplest description, decision tree analysis is a divide and conquer approach for classification. Decision trees can be used to discover features and extract patterns in large databases that are important for discrimination and predictive modelling [21]. A decision tree consists of nodes at which a variable is tested. A variable can be a nominal or numerical value and in the latter case, the test usually determines whether the variable's value is greater or less than a predetermined constant, resulting in a two-way split. A variable is selected to split the data set at the first node (root node). For each possible test outcome at the node, a branch is made ending in a daughter node. The process can be repeated recursively for each branch, using only those records that actually reach the branch. If at any time all records at a node have the same classification, that part of the tree stops developing [18]. The same procedure as for other techniques was done and the accuracy of the model was calculated.

### 2.4.4 | Support vector machine

Support vector machine (SVM) is a supervised learning algorithm that is well suited for determining patterns in complex data sets. It performs the classification by finding a hyperplane that maximizes the margin between classes. The vector determines the hyperplane is considered the support vector. The algorithm performs the classification and learns from the examples to predict the classification of never-before-seen data [22]. To do so, two inputs are needed: training data set and test data set. The class label file clarifies each training example, in this case, each set of samples is represented by a specific number. The goal of model selection is to adjust the hyperparameters (penalty parameters and any kernel parameters) of the SVM classification to achieve the lowest test error, such as the lowest probability of misclassification from unseen test examples [23]. As predictor variables the intensity values depending on the wavelength were utilized, as the response values, each group of samples were put equal to a specific number. The cubic function was implemented as the kernel function. The classification learner was used for the model calculation, for prediction the function predict was applied. Binary classification is employed for classification tasks with two classes while multi-class classification is implemented for classification tasks with more than two classes. Heuristic methods can be used to split a multi-class classification problem into multiple binary classification datasets and train a binary classification model each. One-vs-All is a heuristic method for using binary classification algorithms for multi-class classification whereas the One-vs-One strategy splits a multi-class classification into one binary classification problem per each pair of classes. It involves splitting the multi-class dataset into multiple binary classification problems. In this case, all the strategies were tested to find the best algorithms and finally, One-vs-All was utilized.

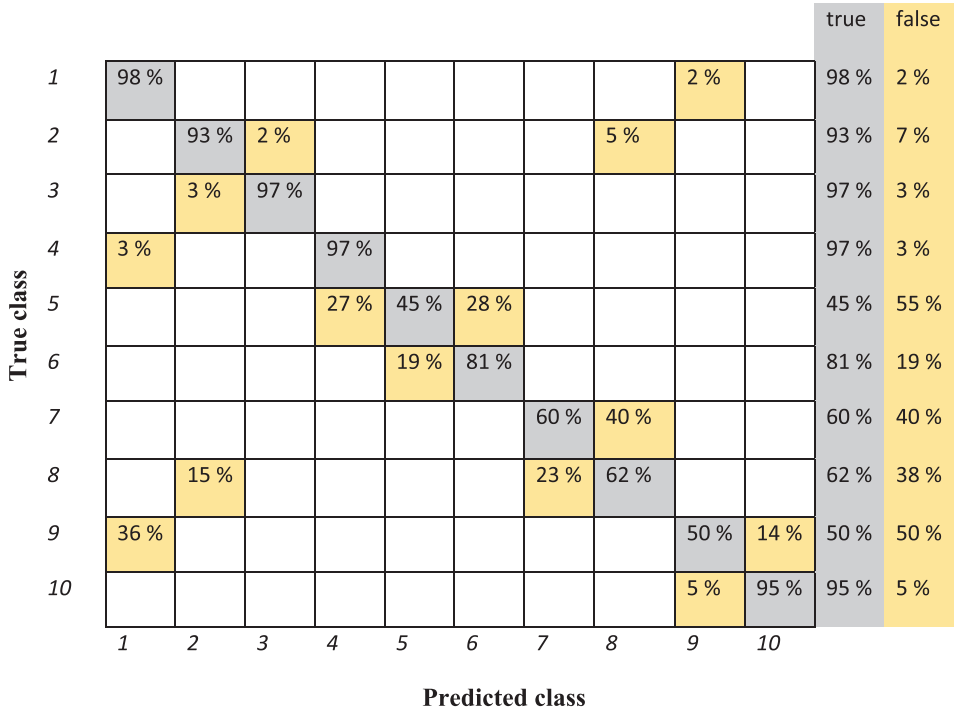


**FIGURE 1** Confusion matrix of classification of various samples from training set by means of linear discriminant analysis. 1: Reference sample, 2: 10% water, 3: 5% water, 4: 1.6% fat, 5: 1.7% fat, 6: 1.8% fat, 7: 5% cleaning solution, 8: 10% cleaning solution, 9: 15°C, 10: 20°C (misclassifications are shown in yellow)

### 2.4.5 | K nearest neighbor

The k nearest neighbor is one of the simplest machine learning algorithms, based on the fact that objects close to each other will show similar characteristics. Therefore, if the characteristics of a sample are obvious, it is easy to predict the characteristics of its neighbors. k is a positive small integer which indicates how many neighbors are considered. The k nearest neighbors are selected based on distance metric and here, the Euclidean was employed [23]. In this contribution, for the training set, each class was named by a number and k was equal to 3. In MATLAB, the classification learner was used for the model calculation, for prediction the function predict was applied.

## 3 | RESULTS AND DISCUSSION

### 3.1 | Linear discriminant analysis

Linear discriminant analysis shows an accuracy of 89.8% for the training set and 67.8% for the test set. Figure 1 represents the result of using linear discriminant analysis for classification in a confusion matrix.

As can be seen in Figure 1, 98% of the reference samples are classified correctly in group one. However, 2% of the mentioned samples are classified in the group of samples which were measured at 15°C (false negative). In contrast, 3% of the samples with 1.6% fat and 36% of samples measured at 15°C are wrongly categorized as reference samples (false positive). The highest wrong classification is related to sample number 7 which is named 5% cleaning solution in which 40% of spectra are classified as the samples with 10% cleaning solution. Although, while 45% of samples with 1.7% fat are correctly classified, 27% of this sample classified as 1.6% fat and 28% as 1.8% fat. By a closer look at Figure 1, just 62% of the sample with 10% cleaning solution classified correctly. The mentioned sample is classified into wrong groups, 23% in the group of 5% cleaning solution and 15% in the group of 10% water. As the training model did not work well, the test set is not discussed here.

### 3.2 | Decision tree

The decision tree shows the accuracy of 83.9% for the training set and 61% for the test set. The confusion matrix obtained by using the decision tree and the training set for classification is presented in Figure 2.

According to Figure 2, there are a lot of misclassifications where each sample is classified into several irrelevant groups. Reference samples are classified as 10% water, 1.6

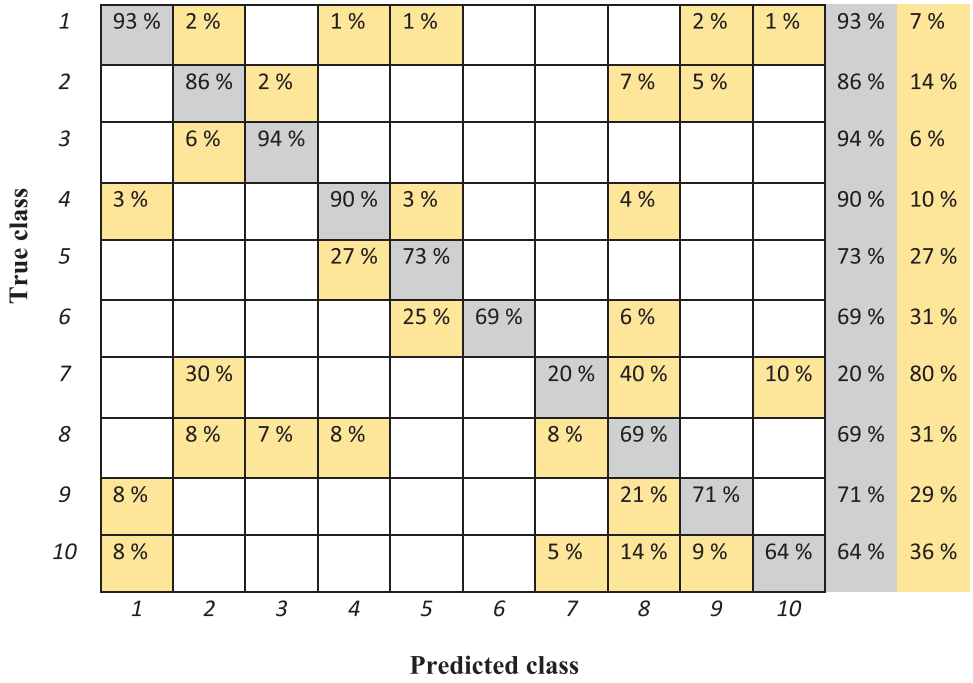


**FIGURE 2** Confusion matrix of classification of various samples from training set by means of decision tree. 1: Reference sample, 2: 10% water, 3: 5% water, 4: 1.6% fat, 5: 1.7% fat, 6: 1.8% fat, 7: 5% cleaning solution, 8: 10% cleaning solution, 9: 15°C, 10: 20°C (misclassifications are shown in yellow)

and 1.7% fat and samples measured at 15°C and 20°C. The best classification is referred to sample with 5% water, of which 94% are classified truly and 6% are categorized as 10% water. The worst classified sample is referred to the sample with 5% cleaning solution, while just 20% of spectra are classified correctly. Forty percent of mentioned samples are in the group of 10% cleaning solution, 30% in the group of 10% water, and 10% in the group of samples measured at 20°C.

## 3.3 | Support vector machine

Figure 3 demonstrates the result of using a support vector machine to classify the training set in the confusion matrix.

The accuracy of the model is 96% for the training set and 81.4% for the test set. According to Figure 3, all the reference samples are classified correctly in group number one. While all the samples with 5% water are classified correctly, 2% of samples with 10% water are classified as 5% water. Three percent of the sample with 1.6% fat and 9% of samples with 1.7% fat are classified wrongly as reference sample.

As shown in Table 2, while most of the sample represents the high value for recall, the sample with 10% water shows a value of just 11%. By accurate investigation, it was found that this sample is mainly classified as 5% water and minimally as 10% cleaning solution. The value of recall for the sample with 10% cleaning solution is 67% when some of the samples are classified as 10% water. The calculated specificities of samples are quite high, while the lowest one is equal to 93% for milks measured at 20°C. The computed precision values for samples with various fat content are 100% implies the ability of this method for correct categorizing of fat content. The lowest precision values are 50% and 55% for the samples with 10% water and measured at 20°C, respectively.

**TABLE 2** Calculated recall, specificity, and precision of the test set for each sample using support vector machine

| Sample | Recall (sensitivity) | Specificity | Precision |
|---|---|---|---|
| Reference sample | 100% | 98% | 93% |
| 10% water | 11% | 98% | 50% |
| 5% water | 100% | 94% | 67% |
| 1.6% fat | 80% | 96% | 67% |
| 1.7% fat | 100% | 100% | 100% |
| 1.8% fat | 80% | 100% | 100% |
| 10% cleaning solution | 67% | 100% | 100% |
| 5% cleaning solution | 75% | 100% | 100% |
| 15°C | 100% | 100% | 100% |
| 20°C | 100% | 93% | 55% |

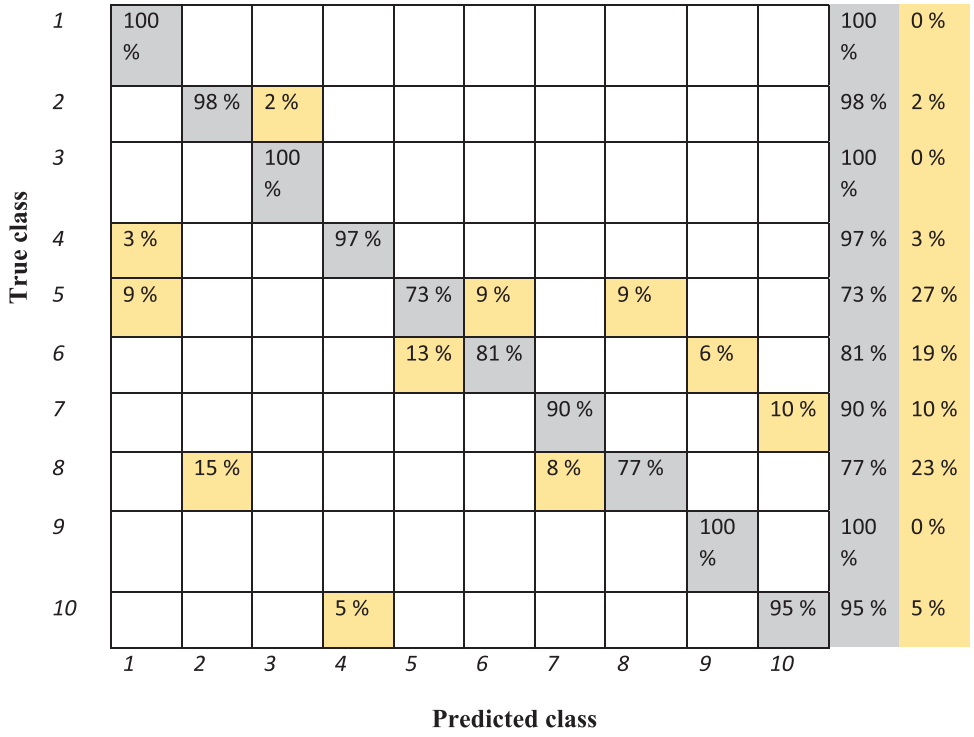


**FIGURE 3** Confusion matrix of classification of various samples from training set by means of support vector machine. 1: Reference sample, 2: 10% water, 3: 5% water, 4: 1.6% fat, 5: 1.7% fat, 6: 1.8% fat, 7: 5% cleaning solution, 8: 10% cleaning solution, 9: 15°C, 10: 20°C (misclassifications are shown in yellow)

## 3.4 | K nearest neighbor

The result of classification of the training set by using k nearest neighbor in the confusion matrix is presented in Figure 4.

The accuracy of the model is 96.4% for the training set and 84.8% for the test set. According to Figure 4, all the reference samples are classified correctly; however, 3% of 1.6% fat and 9% of 1.7% fat are categorized as the reference samples (the same as support vector machine). In comparison with other samples, 5% cleaning solution is not classified as good as others. Ten percent of the mentioned sample is classified wrongly as 1.7% fat and 10% as the samples with 10% cleaning solution. Six percent of the sample with 1.8% fat is categorized wrongly as 1.7% fat and also 6% as 1.6% fat. Samples measured at 20°C are classified quite good; however, 4% and 5% of the spectra are classified as 10% water or 15°C, respectively. Low numbers of spectra related to 10% cleaning solution are classified as 10% water and 1.6% fat wrongly.

As can be seen in Table 3, recall values for the reference sample, 5% water and cleaning solution, 1.7% fat, and 20°C are equal to 100%. Although, the lowest value is contributed to the 10% water which is mostly classified as the 5% water wrongly. The recall value for 10% cleaning solution is 67%, where this sample is mostly categorized as 10% water. In addition, the mentioned value for samples with 1.8% fat content is equal to 75%. The calculated specificity of samples is quite high, while the lowest value is 92% contributed to 5% water in the milk. Despite a hundred per cent precision for most of the samples, samples with 5 and 10% water show the precision of 62% and 60%, respectively. The

**TABLE 3** Calculated recall, specificity, and precision of test set for each sample using k nearest neighbor

| Sample | Recall (sensitivity) | Specificity | Precision |
|---|---|---|---|
| Reference sample | 100% | 98% | 93% |
| 10% water | 56% | 94% | 62% |
| 5% water | 100% | 92% | 60% |
| 1.6% fat | 80% | 100% | 100% |
| 1.7% fat | 100% | 100% | 100% |
| 1.8% fat | 75% | 100% | 100% |
| 10% cleaning solution | 67% | 100% | 100% |
| 5% cleaning solution | 100% | 100% | 100% |
| 15°C | 80% | 100% | 100% |
| 20°C | 100% | 98% | 83% |

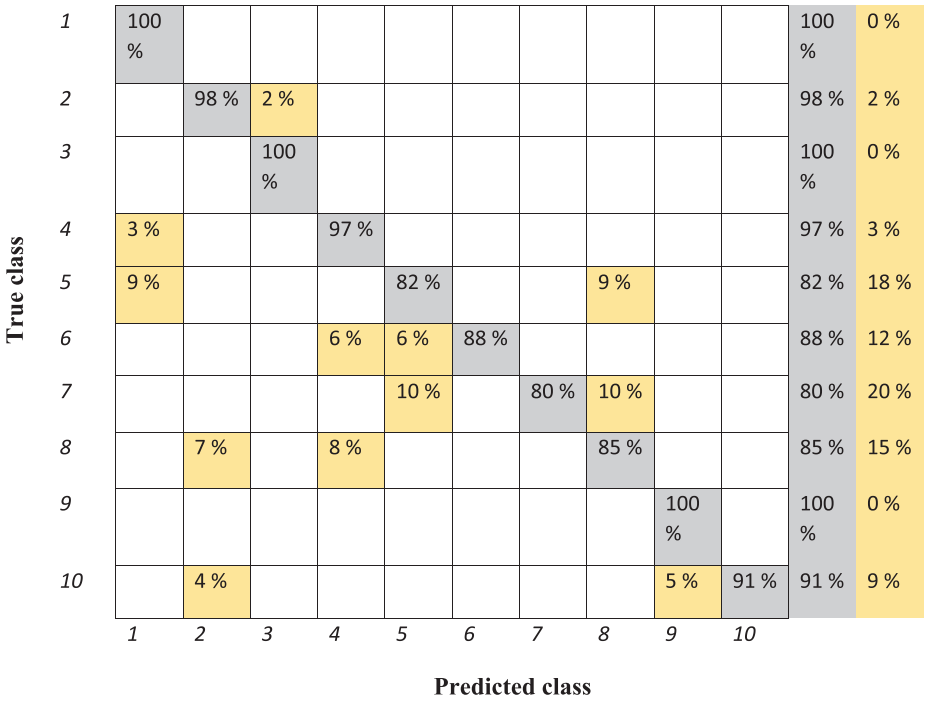


**FIGURE 4** Confusion matrix of classification of various samples from training set by means of k nearest neighbor. 1: Reference sample, 2: 10% water, 3: 5% water, 4: 1.6% fat, 5: 1.7% fat, 6: 1.8% fat, 7: 5% cleaning solution, 8: 10% cleaning solution, 9: 15°C, 10: 20°C (misclassifications are shown in yellow)

precision value for samples measured at 20°C is 83% and for the reference sample is equal to 93%.

## 4 | CONCLUDING REMARKS

The use of a process analyzer to monitor milk processing helps suppliers to maintain product quality and safety before filling and packing. In this research, Raman spectroscopy was used for developing an innovative approach that can detect anomalies from reference processing in a by-pass to reduce the probability of quality-defect recalls. To determine the best machine learning methodology for classifying various anomalies, a variety of classification methods—linear discriminant analysis, decision tree, support vector machine, and k nearest neighbors—were used. The results demonstrate that decision tree and linear discriminant analysis models, with the accuracy of 61% and 67.8% for the test set, respectively, are unable to correctly predict the classes. Also, by taking a look at the confusion matrix of the training set, it is clear that these two methods classify the reference samples as the modified samples and vice versa. Consequently, continuing with them was not fruitful in this application. In contrast, the support vector machine and k nearest neighbor, perform well in the categorization of diverse groups, with the accuracy of 81.4% and 84.8% for the test set, respectively. In this case, the most important thing is that anomalies can be separated from the reference signals. Classifying the various anomalies would be good but not essential. Therefore, it would be necessary to distinguish samples with abnormal changes from the reference sample which support vector machine and k nearest neighbor did. Both procedures show high values of recall, specificity, and accuracy for the reference sample (100%, 98%, and 93%, respectively), indicating that these methods are capable of classification. In addition, these values are quite high for the modified samples imply the fact that most of the abnormal signals are classified correctly by these methods. Therefore, these two methods might work as well for the classification of other spectroscopic applications.

In general, it can be stated that support vector machine and k nearest neighbor are capable of accurately detecting and identifying various anomalies during milk processing, allowing the milk industry to respond quickly to the situation.

### ACKNOWLEDGMENTS

This IGF project of the FEI is/was supported via AiF (20200N) within the programme for promoting the Industrial Collective Research (IGF) of the German Ministry of Economics and Energy (BMWi), bases on a resolution of the German Parliament for which the authors are grateful.

Open access funding enabled and organized by Projekt DEAL.

## CONFLICT OF INTEREST

The authors have declared no conflict of interest.

## DATA AVAILIBILTY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

## REFERENCES

1. Teneva-Angelova, T., Balabanova, T., Boyanova, P. and Beshkova, D., Traditional Balkan fermented milk products. Eng. Life Sci. 2018, 18, 807–819.
2. Úbeda, M. A., Hussein, W. B., Hussein, M. A., Hinrichs, J., et al. Acoustic sensing and signal processing techniques for monitoring milk fouling cleaning operations. Eng. Life Sci. 2016, 16, 67–77.
3. Vasafi, P. S., Paquet-Durand, O., Brettschneider, K., Hinrichs, J., et al. Anomaly detection during milk processing by autoencoder neural network based on near-infrared spectroscopy. J. Food Eng. 2021, 299, 110510.
4. Wallhäußer, E., Hussein, W. B., Hussein, M. A., Hinrichs, J., et al. Detection of dairy fouling: Combining ultrasonic measurements and classification methods. Eng. Life Sci. 2013, 13, 292–301.
5. Cheng, Y., Dong, Y., Wu, J., Yang, X., et al. Screening melamine adulterant in milk powder with laser Raman spectrometry. J. Food Compos. Anal. 2010, 23, 199–202.
6. McGoverin, C. M., Clark, A. S. S., Holroyd, S. E. and Gordon, K. C., Raman spectroscopic quantification of milk powder constituents. Anal. Chim. Acta 2010, 673, 26–32.
7. Reiner, J., Protte, K. and Hinrichs, J., Investigation of the applicability of Raman spectroscopy as online process control during Consumer milk Production. ChemEngineering 2020, 4, 45.
8. Ciosek, P., Brudzewski, K. and Wróblewski, W., Milk classification by means of an electronic tongue and support vector machine neural network. Meas. Sci. Technol. 2006, 17, 1379.
9. Borin, A., Ferrao, M. F., Mello, C., Maretto, D. A., et al. Least-squares support vector machines and near infrared spectroscopy for quantification of common adulterants in powdered milk. Anal. Chim. Acta 2006, 579, 25–32.
10. Lianou, A., Malavazos, C., Triantafyllou, I., George-John, E. N., et al. Rapid assessment of the microbiological quality of pasteurized vanilla cream by means of Fourier transform infrared spectroscopy in tandem with support vector machine analysis. Food Anal. Methods 2018, 11, 840–847.
11. Amjad, A., Ullah, R., Khan, S., Bilal, M., et al. Raman spectroscopy-based analysis of milk using random forest classification. Vib. Spectrosc. 2018, 99, 124–129.
12. de Lima, G. F., Andrade, S. A. C., da Silva, V. H. and Honorato, F. A., Multivariate classification of UHT milk as to the presence of lactose using benchtop and portable NIR spectrometers. Food Anal. Methods 2018, 11, 2699–2706.
13. Santos, P. M., Pereira-Filho, E. R. and Colnago, L. A., Detection and quantification of milk adulteration using time domain nuclear magnetic resonance (TD-NMR). Microchem. J. 2016, 124, 15–19.
14. Steensels, M., Antler, A., Bahr, C., Berckmans, D., et al. A decision-tree model to detect post-calving diseases based on rumination, activity, milk yield, BW and voluntary visits to the milking robot. Animal 2016, 10, 1493–1500.
15. Vasafi, P. S., Hinrichs, J. and Hitzmann, B., Establishing a novel procedure to detect deviations from standard milk processing by using online Raman spectroscopy. Food Control 2021, 108442.
16. Lasch, P., Spectral pre-processing for biomedical vibrational spectroscopy and micro spectroscopic imaging. Chemom. Intell. Lab. Syst. 2012, 117, 100–114.
17. Rinnan, Å., Van Den Berg, F. and Engelsen, S. B., Review of the most common pre-processing techniques for near infrared spectra. TrAC Trends Anal. Chem. 2009, 28, 1201–1222.
18. Lazar, A. and Shellito, B., The classification of imbalanced spatial data. In MAICS, 2011. pp. 108–113.
19. Fisher, RA. The use of multiple measurements in taxonomic problems. Ann Eugenic 1936, 7, 179–188.
20. Ricciardi, C., Valente, A. S., Edmund, K., Cantoni, V., et al. Linear discriminant analysis and principal component analysis to predict coronary artery disease. Health informatics journal 2020, 26, 2181–2192.
21. Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A. et al„ An introduction to decision tree modeling. Journal of Chemometrics: A Journal of the Chemometrics Society 2004, 18, 275–285.
22. Gold, C. and Sollich, P., Model selection for support vector machine classification. Neurocomputing 2003, 55, 221–249.
23. Khamis, H. S., Cheruiyot, K. W. and Kimani, S., Application of k nearest neighbor classification in medical data mining. International Journal of Information and Communication Technology Research 2014, 4.

**How to cite this article:** Vasafi, P. S. and Hitzmann, B. Comparison of various classification techniques for supervision of milk processing. Eng Life Sci. 2021;1–9.
https://doi.org/10.1002/elsc.202100098

# 4. Discussion

Milk and dairy products are well-known for their health-promoting properties. Milk is a worldwide source of protein, lactose, vitamins, minerals, and lipids and is one of the most important components of the human diet (Teneva-Angelova, 2018). Large resource investments are required to reduce the limiting difficulties arising from food industrial intensive productivity (Beda et al., 2016). Dairy producers benefit from using rapid spectroscopy to detect product flaws online since it allows them to change product features or redirect product flow during the manufacturing process. During the processing of milk, some abnormal changes – fat and temperature variation, fouling, added water and cleaning solution - can happen which threaten the quality and safety of final products. Because the fat content of milk products is usually fixed during processing, the amount of fat used during manufacture must be consistent and accurate. A pilot plant is usually cleaned with water and cleaning solution, but its presence in the production line raises various safety and quality problems. As milk is a heat-treated product, temperature is the most crucial factor to consider. As a result, managing these changes is a worthwhile endeavor (Vasafi et al, 2021). Food quality and food safety have been a popular topic in the media as a result of recent so-called food crises in Europe. Food quality and food safety are frequently used interchangeably. There are significant disparities, particularly when it comes to food company communication and customer attitudes. Before, most people simply assumed that all food on the market matched these two criteria. This was self-evident, and the consumer didn't need to be informed about food safety. Food safety has become a food quality attribute in recent years, which has changed the situation. The food and feed industries are being pushed by government officials to build comprehensive quality management systems to improve food safety, restructure the food inspection system, and improve consumer information in order to recover consumer faith in food. Food producers are required to monitor information relating to consumer perceptions of whether the organization has met customer needs as one of the measurements of the quality management system's performance. Food producers are now interested in communicating food safety because it has become a qualitative trait (Röhr et al., 2005). In order to ensure a satisfying product, the milk processing business has tight requirements on monitoring and control of temperature, composition such as fat concentration, hygiene, taste,

and odor (Vasafi et al., 2021). Therefore, developing an online tool monitors the process can help companies to avoid suffering from production of milks with quality and safety issues.

NIR spectroscopy was used earlier to quantify goat milk adulteration caused by the addition of cow milk, as well as to determine the fat and protein levels. As a result, the proposed methodology appeared to be a helpful, quick, and non-destructive for determining the fat and protein content of goat milk, as well as screening the quality of goat milk for adulteration with cow milk (Dos Santos Pereira et al., 2020). Balabin et al. (2011) illustrated that near infrared spectroscopy is a beneficial tool to detect melamine in dairy products, such as infant formula, milk powder, or liquid milk. This method can be considered as a fast, robust, sensitive, and low-cost method for dairy industry.

Within this thesis, anomaly detection during milk processing by using spectroscopy was investigated. By using NIR spectroscopy, 470 spectra were taken from different packages of UHT-milk with 1.5 % of fat from the brand "Weihenstephan," Germany and used as the training set. For the prediction set, different samples were measured. Various samples such as milk with 1.5 % fat, both ESL-milk (extended shelf life) and UHT-milk from various suppliers were measured to check if they show any detectable differences from each other. Same samples used for training set were measured after heating up to 10 °C or 15 °C to determine the sensitivity of samples to the temperature. Sample with 1.6 % fat was prepared by mixing 95 mL of 1.5 % fat milk and 5 mL of 3.5 % fat milk to define if difference of 0.1 % fat content can be distinguished. Different concentrations of water or cleaning solution were added to the reference milk and measured after cooling down to 5 °C to understand how much water and cleaning solution can be detected in the milk. A cold-water bath with temperature of 5 °C was implemented to keep the measurement temperature stable. The flow cell was connected by plastic tubes to the source of milk in the beaker. Milk was circulated through the flow cell during the measurement for two reasons. First, in order to keep the temperature stable and and second, to simulate on-line process measurement conditions in a by-pass. The approach of using spectroscopy techniques to monitor milk processing seems to be fruitful in order to help companies to produce high quality products and also to avoid suffering. The combination of using near infrared spectroscopy (NIR) and an autoencoder neural network seems to be a useful procedure to detect anomalies. Here the ability of an autoencoder to be used as a self-learning tool has many advantages, such as no further experimental effort or offline measurements. However, various pre-processing methods presented different results. Using

first derivative spectra, the trained autoencoder was able to detect 5 % additional water and 9 % cleaning solution in the milk. It was also able to detect a 0.1 % change in fat contents using the combination spectra. Furthermore, both procedures were able to detect changes in milks from various suppliers as well as a temperature variation of 10 °C. The results were promising demonstrating the ability to use an autoencoder as self-learning procedure. In NIR measurements, characteristic absorption bands of milk components can be affected by the high absorption by water in combination with the strong light scattering by the fat globules. Therefore, an alternative measurement system might help to improve the self-learning capability of such a system.

Raman spectroscopy was applied earlier for fast determination of milk fat content and considered as a quick beneficial technique which can reveal contributions from proteins, but mainly from their fat content with different spectral characteristics (El-Abassy et at., 2011). Raman scattering, like MIR absorption spectroscopy, provides access to the vibrational fingerprints of molecules. Raman spectroscopy has recently been used in food analysis due to its ability to offer content-relevant information based on well-defined and resolved spectra in a variety of sample categories, including liquids, solids, and gases. Raman spectroscopy is based on inelastic light scattering, a process in which incident excitation radiation interacts with molecules and causes a red shift in the energy of the incident photon equal to the internal vibrational energies (El-Abassy et at., 2011). As the disadvantages of NIR for milk analysis with high amount of water and fat globules were explained, Raman spectroscopy were utilized in order to improve the results of the self-learning system.

In this case, circulation method was utilized again and spectra were taken from both reference and modified samples. For Raman spectroscopy, reference sample with 1.5 % UHT fat was measured at 10 °C, as the heat of laser affected the temperature of samples. Samples with various changes such as different fat content, various measurement temperature (10 °C and 15 °C), and added water and cleaning solution were measured by Raman spectroscopy as the modified samples. Development of an autoencoder neural network based on Raman spectra represented better results which were able to detect 5 % water or cleaning solution in the milk. Also, a difference of 0.1 % in fat content and temperature variation of 5 °C were detected by this procedure. Moreover, other procedures such as principal component analysis and regression, partial least squares regression, and Gaussian process regression were developed to check the possibility of change detection and their ability as evaluation methods in a self-

learning module. Principal components analysis (PCA) classified abnormal signals as normal signals and was not able to classify anomalies correctly. Principal component regression (PCR) also was not capable to predict abnormal and normal signals accurately while it presented the $R^2$ equal to 0.7. In this case, all the normal signals were put equal to zero and modified signals were put equal to one. The model was trained with both reference and modified signals and results presented that a reliable limit of detection which can separate the normal signals from abnormal signals cannot be defined. Meanwhile, Gaussian process regression with $R^2$ equal to 0.97 predicted normal and abnormal signals perfectly. It was capable to detect 5 % water and cleaning solution in milk, 5 °C temperature variation and 0.1 % difference in fat content. Gaussian process regression can detect even more abnormal signals than the autoencoder, but it should be trained with not only reference signals but also all the possible abnormal signals which can be counted as the disadvantage of this method. Therefore the Gaussian process regression increase the workload for the implementation of a self-learning system. The autoencoder can be trained once and just with normal signals and it can be considered as a powerful self-learning system to control milk processing.

Several studies have been carried out for the implementation of classification techniques such as support vector machine to categorize milk samples which showed promising results (Ciosek et al., 2006). In the previous parts just, detection of anomalies was investigated. However, after anomaly detection during the process, it is very important to understand which anomalies or changes happened during the process which can help companies to solve the problem easier and faster. For this reason, several classification methods such as support vector machine, k nearest neighbour, linear discriminant analysis, and decision tree were investigated to get more information out of the spectra. Results illustrated that two procedures called linear discriminant analysis and decision tree showed low accuracy analysing the train set and consequentially were not able to classify abnormal signals and normal ones accurately. On the other hand, support vector machine and k nearest neighbour presented the higher accuracy values not only for the training set but also for the test set. The test set accuracy of 81.4 % for support vector machine and 84.8 % for the k nearest neighbor presented the usefulness of these algorithms for determination of type of anomalies. Both were capable to classify reference and abnormal signals properly which is the most important criterion to select the best classification method. For a comparison, both support vector machine and k nearest neighbour were capable to determine the type of anomalies in the milk

processing, however, based on method accuracy, k nearest neighbour showed better results. For the reference sample, both approaches showed the high value of recall, specificity, and accuracy (100 %, 98 %, and 93 %, respectively), showing that they are capable of categorization.

Overall, the combination of NIR and the autoencoder neural network and utilizing various pre-processing methods was capable to detect difference of 0.1 % in fat content, 10 °C in temperature measurement and milk from various suppliers. In addition, results showed the capability of this method to detect 5 % water and 9 % cleaning solution in milk. However, Raman spectroscopy presented better results than the NIR spectroscopy as it was not affected by high absorption by water and light scattering of fat globules in the milk. The combination of Raman spectroscopy and the autoencoder was able to detect difference of 0.1 % in fat content, 5 °C in temperature of measurement, and 5 % water and cleaning solution in the milk. The combination of Raman spectroscopy and Gaussian process regression presented the same results however, it detected more abnormal signals than the autoencoder. As this method needed to be trained by both normal and abnormal signals, which will increase the effort to be implemented as evaluation procedure in a self-learning module, the autoencoder remained the most proper procedure. It can be trained once and just with normal signals and it is ready for detection. After detection of anomalies in the process, to get more details about the type of anomalies, the classification techniques were really helpful. By using Raman spectroscopy, support vector machine classified anomalies appropriately when it showed the accuracy of 96 % for training set and 81.4 % for test set. By this method, all the reference samples classified correctly in one group. Also using k nearest neighbour, the accuracy of model was 96.4 % for training set and 84.8 % for test set. By this method, all the reference samples classified correctly as well. It can be stated that support vector machine and k nearest neighbor were capable of accurately detecting various anomalies during milk processing. This result can help companies to avoid being suffered by producing lots of samples with quality and safety issues. It can warn the operator when an issue arises, enable them to react immediately. It assists the companies, not only to understand some abnormal changes exist in the process, but also help them to know what is the exact problem. In this work, milk was used as a case of study, however this procedure is able to detect abnormal changes that can happen during the processing of any liquid. In addition even the supervision of powder or any

material, which can be transported in front of the spectroscopic sensor, can be carried out by such a self-learning system.

# Bibliography

Abdi, H., Williams, L. J., 2010. Principal component analysis. Wiley interdisciplinary reviews: *computational statistics*, 2(4), 433-459.

Aye, S. A., Heyns, P. S., 2017. An integrated Gaussian process regression for prediction of remaining useful life of slow speed bearings based on acoustic emission. *Mechanical Systems and Signal Processing*, 84, 485-498.

Balabin, R.M. and Smirnov, S.V., 2011. Melamine detection by mid-and near-infrared (MIR/NIR) spectroscopy: a quick and sensitive method for dairy products analysis including liquid milk, infant formula, and milk powder. *Talanta,* 85(1), pp.562-568.

Bernardo, J., Berger, J., Dawid, A. P. A. F. M. S., Smith, A., 1998. Regression and classification using Gaussian process priors. Bayesian statistics, 6, 475.

Cen, H., He, Y., Huang, M., 2006. Measurements of soluble solids contents and pH in orange juice using chemometrics and Vis–NIRS. *Journal of Agricultural Food Chemistry*, 54, 7437–7443.

Chandan, R.C., 2011. Dairy ingredients for food processing: an overview. *Dairy ingredients for food processing*, pp.3-33.

Chaurasia, P., Younis, K., Qadri, O. S., Srivastava, G., Osama, K., 2019. Comparison of Gaussian process regression, artificial neural network, and response surface methodology modeling approaches for predicting drying time of mosambi (Citrus limetta) peel. *Journal of Food Process Engineering*, *42*(2), e12966.

Chen, T., Morris, J., Martin, E., 2007. Gaussian process regression for multivariate spectroscopic calibration. *Chemometrics and Intelligent Laboratory Systems*, *87*(1), 59-71.

Ciosek, P., Brudzewski, K. and Wróblewski, W., 2006. Milk classification by means of an electronic tongue and support vector machine neural network. *Measurement Science and Technology*, *17*(6), p.1379.

De Luca, A.C., Dholakia, K. Mazilu, M., 2015. Modulated Raman spectroscopy for enhanced cancer diagnosis at the cellular level. *Sensors,* 15(6), pp.13680-13704.

Dijkstra, R. J., Ariese, F., Gooijer, C., Brinkman, U. T. (2005). Raman spectroscopy as a detection method for liquid-separation techniques. *TrAC Trends in Analytical Chemistry*, 24(4), 304–323.

dos Santos Pereira, E.V., de Sousa Fernandes, D.D., de Araújo, M.C.U., Diniz, P.H.G.D., Maciel, M.I.S., 2020. Simultaneous Determination of Goat Milk Adulteration with Cow Milk and Their Fat and Protein Contents Using NIR Spectroscopy and PLS Algorithms. *LWT*, p. 109427.

Gautam, G., Yadav D., 2014. Sentiment analysis of twitter data using machine learning approaches and semantic analysis Contemporary computing (IC3), seventh international conference on, IEEE (2014), pp. 437-442

He, H., Sun, D.W., Pu, H., Chen, L. and Lin, L., 2019. Applications of Raman spectroscopic techniques for quality and safety evaluation of milk: A review of recent developments. *Critical reviews in food science and nutrition*, *59*(5), pp.770-793.

Hou, M., Huang, Y., Ma, L. and Zhang, Z., 2016. Quantitative analysis of single and mix food antiseptics basing on SERS spectra with PLSR method. *Nanoscale research letters*, 11(1), pp.1-8.

Júnior, P.H.R., de Sá Oliveira, K., de Almeida, C.E.R., De Oliveira, L.F.C., Stephani, R., da Silva Pinto, M., de Carvalho, A.F., Perrone, Í.T., 2016. FT-Raman and chemometric tools for rapid determination of quality parameters in milk powder: Classification of samples for the presence of lactose and fraud detection by addition of maltodextrin. *Food Chemistry*, *196*, pp.584-588.

Kawasaki, M., Kawamura, S., Tsukahara, M., Morita, S., Komiya, M., Natsuga, M., 2008. Near-infrared spectroscopic sensing system for on-line milk quality assessment in a milking robot. Computers and electronics in agriculture, 63(1), 22-27.

Krivoshiev, G. P., Chalucova, R. P., Moukarev, M. I., 2000. A possibility for elimination of the interference from the peel in nondestructive determination of the internal quality of fruit and vegetables by vis–NIR spectroscopy. Lebensm-Wiss University of technology, 33, 344–353.

Kulmyrzaev, A.A., Levieux, D. and Dufour, É., 2005. Front-face fluorescence spectroscopy allows the characterization of mild heat treatments applied to milk. Relations with the denaturation of milk proteins. *Journal of Agricultural and Food Chemistry*, 53(3), pp.502-507.

Lohumi, S., Kim, M.S., Qin, J. Cho, B.K., 2017. Raman imaging from microscopy to macroscopy: quality and safety control of biological materials. *TrAC Trends in Analytical Chemistry*, 93, pp.183-198.

Morris, M. D., 2006. Review - modern Raman spectroscopy: A practical approach. *Analytical Chemistry 78*(1):33.

Moros, J., Garrigues, S., de la Guardia, M., 2007. Evaluation of nutritional parameters in infant formulas and powdered milk by Raman spectroscopy. *Analytica chimica acta*, 593(1), pp.30-38.

Næs, T., Martens, H., 1988. Principal component regression in NIR analysis: viewpoints, background details and selection of components. *Journal of chemometrics*, 2(2), 155-167.

Osborne, B. G. 2000. *Near infrared spectroscopy in food analysis* (pp. 1–14). Australia: BRI Australia Ltd.

Qin, J., M. S. Kim, K. Chao, S. Dhakal, H. Lee, B. K. Cho, and C. Mo., 2017. Detection and quantification of adulterants in milk powder using high-throughput Raman chemical imaging technique. *Food Additives & Contaminants* 34(2):152–61.

Raman, C. V., Krishnan, K. S., 1928. A new type of secondary radiation. Nature 121(3048):501-2.

Ranganathan, A., Yang, M. H., Ho, J. 2010. Online sparse Gaussian process regression and its applications. *IEEE Transactions on Image Processing*, *20*(2), 391-404.

Röhr, A., Lüddecke, K., Drusch, S., Müller, M.J. Alvensleben, R.V., 2005. Food quality and safety—consumer perception and public health concern. *Food control*, 16(8), pp.649-655.

Schulz, H., Özkan, G., Baranska, M., Krüger, H. Özcan, M., 2005. Characterisation of essential oil plants from Turkey by IR and Raman spectroscopy. *Vibrational Spectroscopy,* 39(2), pp.249-256.

Sharma, A.K., Sharma, R.K., Kasana, H.S., 2007. Prediction of first lactation 305-day milk yield in Karan Fries dairy cattle using ANN modeling. *Applied Soft Computing,* 7(3), 1112–1120.

Stefanov, I., Baeten, V., Abbas, O., Colman, E., Vlaeminck, B., De Baets, B. Fievez, V., 2010. Analysis of milk odd-and branched-chain fatty acids using Fourier transform (FT)-Raman spectroscopy. *Journal of agricultural and food chemistry*, 58(20), pp.10804-10811.

Teneva-Angelova, T., Balabanova, T., Boyanova, P. and Beshkova, D., 2018. Traditional Balkan fermented milk products. *Engineering in Life Sciences*, *18*(11), pp.807-819.

Tripathy, A., Agrawal, A., Rath, S.K., 2016. Classification of sentiment reviews using n-gram machine learning approach. *Expert Systems with Applications*, *57*, pp.117-126.

Varnam, A. H., Sutherland, J. P., 1994. Milk and milk products: Technology, chemistry and microbiology. *International Dairy Journal* 5(7):721–7.

Vasafi, P.S., Paquet-Durand, O., Brettschneider, K., Hinrichs, J., Hitzmann, B., 2021. Anomaly detection during milk processing by autoencoder neural network based on near-infrared spectroscopy. *Journal of Food Engineering*, 299, p.110510.

Yang, D. and Ying, Y., 2011. Applications of Raman spectroscopy in agricultural products and food analysis: A review. *Applied Spectroscopy Reviews*, 46(7), pp.539-560.

Zhang, L. and Henson, M.J., 2007. A practical algorithm to remove cosmic spikes in Raman imaging data for pharmaceutical applications. *Applied spectroscopy*, 61(9), pp.1015-1020.

**Annex**

**Declaration in lieu of an oath on independent work**

**according to Sec. 18(3) sentence 5 of the University of Hohenheim's Doctoral Regulations for the Faculties of Agricultural Sciences, Natural Sciences, and Business, Economics and Social Sciences**

1. The dissertation submitted on the topic

Self-learning modules for spectra evaluation

is work done independently by me.

2. I only used the sources and aids listed and did not make use of any impermissible assistance from third parties. In particular, I marked all content taken word-for-word or paraphrased from other works.

3. I did not use the assistance of a commercial doctoral placement or advising agency.

4. I am aware of the importance of the declaration in lieu of oath and the criminal consequences of false or incomplete declarations in lieu of oath.

I confirm that the declaration above is correct. I declare in lieu of oath that I have declared only the truth to the best of my knowledge and have not omitted anything.

09.09.2021 , Stuttgart — Pegah Saeleghi

Place, Date — Signature

**Pegah Sadeghi Vasafi**

**Date of birth: July 14,1992**

**sadeghy.pegah@yahoo.com (+49) 17655059706**

**Address: Echterdinger str.12, 70599, Stuttgart, Germany**

Skilled PhD candidate in the field of Natural Science, experienced in developing new data-driven tools for online monitoring and supervision of milk processing and employing spectroscopic methods for quality control.

## Work experience

**2019-Present** **Researcher,** University of Hohenheim, Stuttgart, Germany
Project: Development of an online monitoring system for the early detection of product instabilities

- Developed novel analytical methods based on spectroscopy
- Evaluated large dataset acquired by various spectrometers
- Implemented a neural network to detect anomalies
- Gained experience in machine learning and deep learning algorithms
- Performed regression analysis and predictive models to define the problem in the case of downtime
- Utilized classification techniques such as support vector machine, k nearest neighbor, and decision tree

**2017-2018** **Researcher,** University of Hohenheim, Isfahan, Iran

- Determined the best time and temperature for baking of ''Barbari'' bread partially
- Analyzed the effects of freezing storage on the texture, color, volume, and microbial growth of part-baked bread
- Analyzed the effects of modified atmosphere packaging on quality and microbial stability of part-baked bread

## Education

**2019-Present** **PhD in Natural Science**
University of Hohenheim, Stuttgart, Germany
Thesis: Self-learning modules for spectra evaluations
Supervisors: Prof. Dr. Bernd Hitzmann, Prof. Dr. Jörg Hinrichs

**2015-2017** **Master of Food Engineering**
Isfahan University of Technology, Isfahan, Iran (Grade: 16.92 out of 20[1])

**2010-2014** **Bachelor of Food Science and Technology**
University of Jahrom, Fars, Iran (Grade: 17 out of 20)

[1] 20 is considered as the best mark and 1 as the lowest mark.

**Publications**

**2021**

- Comparison of various classification techniques for detection of anomalies during the milk processing, Journal of Engineering in Life Sciences, Wiley Online Library
- Deviations from standard milk processing by using online Raman spectroscopy, Journal of Food Control, Elsevier
- Anomaly detection during milk processing by autoencoder neural network based on near-infrared spectroscopy, Journal of Food Engineering, Elsevier

**2019**

- Quality and microbial stability of part-baked 'Barbari bread' during freezing storage, LWT, Elsevier

**Congress**

**2021** Lecturer at DECHEMA ,13th European Congress of Chemical Engineering and 6th European Congress of Applied Biotechnology, Germany

**Workshops**

**2014** HACCP workshop, Isfahan, Iran
GMPs (Good Manufacturing Practices) workshop, Isfahan, Iran

**Skills**

Good knowledge of MATLAB
Good knowledge of statistical applications such as Spss, Statistix
Good knowledge of Unscrambler and Image J
Good knowledge of Microsoft office packages
Good communicational and organizational skills

**Languages** English (Excellent command)
German (Intermediate command, B2 _ still learning)
Persian (native)

**Interests** Travelling, reading, running, volleyball

Stuttgart 9.9.2021 Pegah Saadeghi