Measurement 258 (2026) 119246

Contents lists available at ScienceDirect

Measurement

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

On the applicability of speckle pattern imaging combined with AI for raw milk classification

Cristina Nuzzi [a,*], Simone Pasinetti [a], Irene Bassi [b], Valentina Bello [b]

[a] University of Brescia, Department of Mechanical and Industrial Engineering, Via Branze 38, Brescia, 25123, Italy
[b] University of Pavia, Department of Electrical, Computer and Biomedical Engineering, Via Adolfo Ferrata 5, Pavia, 27100, Italy

ARTICLE INFO

Keywords:
Measurement science
Speckle pattern imaging
Uncertainty
Raw milk analysis
Artificial intelligence

ABSTRACT

This work demonstrates the high potential of an innovative technique for raw milk classification based on the integration of speckle pattern imaging and artificial intelligence. By exciting speckle patterns with a semiconductor laser and collecting experimental images with a CMOS camera, a total of 20 samples of raw cow milk with similar nutritional contents were tested during 4 Campaigns. Data analysis was conducted leveraging one common feature-based machine learning model and one state-of-the-art image-based deep learning model for speckle patterns. This study aims to provide in-depth insights to the community on how this measurement technique can be applied to raw cow milk samples and how the prediction models tested perform due to the similarity of the nutritional components of the samples. The machine learning model was trained on a set of 16 custom features, while the deep learning model used speckle pattern images as input. Both types of data were standardized dataset-wise beforehand using z-score. The best machine learning and deep learning models achieved 95% accuracy. The study highlights that the nutritional similarity of the samples highly impacts the models' confusion in both cases, especially when Campaigns conducted at different sample temperatures were not included in the training. Overall, the analysis technique presented leveraging uncertainty metrics is a stepping stone toward relevant advances in the field of milk analysis.

1. Introduction

Milk is often defined as the "perfect food": it is rich in fundamental nutrients such as carbohydrates, proteins, lipids, minerals, and vitamins. For this reason, milk is consumed daily by billions of people of all ages worldwide, including millions of infants and children. It also serves as the raw material for many dairy derivatives, such as cheese, yogurt, and butter. In 2024, roughly 555 million tons of cow milk and 120 million tons of other animal milks were produced worldwide [1]. In particular, India is confirmed to be the largest milk producer with a total of 212 million tons of milk, followed by European Union (150 million tons), USA (102 million tons), and China (44 million tons). In Europe, the major cow milk producers are Germany, France, the Netherlands, Italy, and Poland, accounting for 22%, 16%, 10%, 9%, and 9% of the total European production, respectively [2]. The average world price of farmgate milk has been around 0.50 €/kg, with an increasing trend over the last twenty years, yielding to a market business of hundreds of billion euros [1].

Determining raw milk composition is crucial for many reasons. First of all, it allows to establish its quality in terms of nutritional content. Moreover, it is fundamental for farmers to select the most suitable breeds, adjust the dietary plan of the mammals, and understand how it affects milk properties. Last, milk composition has also an impact on the final quality and characteristics of dairy derivatives. The composition of raw milk (i.e. milk just collected from the animals and not yet subject to any type of further processing) is affected by several factors, such as the genotype of the cows, the lactation stage, the dietary regime [3], and even environmental conditions (e.g. climate and season) [4].

Currently, milk composition and properties are measured in dedicated laboratories with very complex techniques [5–8]. Milk samples need to be collected from production sites and then sent to external laboratories where fluids are typically manipulated and pre-treated. Chemical tests are the most commonly used to assess milk composition. For example, the acido-butyrometric method (well-known as Gerber method) is a reference technique for determining the fat content of milk, as the basis of the ISO 19662 and IDF 238:2018 standards [9]. Lipids are separated from proteins by centrifugation, by adding sulfuric acid amyl alcohol. The Röse–Gottlieb procedure is another standard method (ISO 23318:2022, IDF 249:2022) for lipid quantification in

* Corresponding author.
E-mail addresses: cristina.nuzzi@unibs.it (C. Nuzzi), simone.pasinetti@unibs.it (S. Pasinetti), irene.bassi@unipv.it (I. Bassi), valentina.bello@unipv.it (V. Bello).

https://doi.org/10.1016/j.measurement.2025.119246
Received 19 May 2025; Received in revised form 15 September 2025; Accepted 4 October 2025

milk [10], based on a complex chemical/gravimetric procedure: it involves a dual organic solvent system (diethyl ether and petroleum ether) and the addition of a strong mineral alkali in combination with heat to dissociate the lipid–protein complexes. The Kjeldahl method is the reference approach (ISO 1871:2009) for the quantification of protein [11]: it uses digestion, distillation, and titration to arrive at the nitrogen content of the sample. High-performance liquid chromatography is the reference method to determine lactose content in raw milk (ISO 22662:2024) [12]. These methods are considered the gold standards but require specialized laboratories and personnel, involve very long preparation and pre-treatment of the sample with toxic and dangerous chemicals, and are expensive and time-consuming. Hence, they have several drawbacks and they are not suitable for frequent on-site testing of milk samples.

Another more recent method used to measure constituents of milk in laboratories is based on mid-infrared (MIR) spectroscopy. For example, Soyeurt et al. proposed MIR spectroscopy to estimate fatty acid content in cow milk as an alternative to gas–liquid chromatography [13]. In [14], Etzion et al. investigated MIR attenuated total reflectance spectroscopy to determine protein concentration in raw cow milk, exploiting two absorbance bands (1500–1700 cm^{-1} and 1060–1100 cm^{-1}). Since milk is a highly diffusive medium, and the light penetration depth is limited, MIR spectra are very sensitive to the generation of fat biofilms that may form on the walls of cuvettes or pipes, MIR spectroscopy is not suitable for online raw milk analysis [15]. Homogenization could improve measurement accuracy, but it is not practicable for on-line analysis. Moreover, MIR spectra need careful data polishing, fitting statistics, and calibration to extract reliable information [16]. An alternative technique is near-infrared (NIR) spectroscopy [17–20], which is cheaper than MIR spectroscopy, although less accurate. For example, in [21], the authors developed a NIR spectroscopic sensing system for assessing milk quality during robotic milking by collecting and analyzing transmittance spectra through raw milk in the wavelength range 600–1050 nm. In another work [22], the milk absorption spectra from 1100 to 2400 nm were obtained with a spectrophotometer using a 1-mm-thick quartz cuvette and by processing spectral data with a Fourier transform. The content of fat, total protein, and lactose content of non-homogenized milk was determined. Many commercial devices for milk testing, such as MilkoScan (FOSS Analytical, Denmark), DairySpec FT (Bentley Instruments, MN, USA), and Milk Analyzers (Ekomilk Horizon, Bulgaria), are based on MIR/NIR technologies; however, they cost several thousand euros. Enzymatic reaction-based techniques, biosensors, and gas sensors are also widely exploited in milk analyses for detecting lactose [23,24], studying proteins [25,26], monitoring spoilage [27,28], and detecting cow mastitis and other diseases that compromise milk quality [29,30]. Like chemical tests, these methods often require sample preparation, for which chemicals and laboratory tools are necessary, and cannot be performed in real-time.

Among optical sensing techniques, speckle pattern (SP) imaging is surely one of the most innovative and worthy to be investigated. SP analyses can be performed in a contactless and non-invasive manner, even with low-cost instrumentation, and they do not require the treatment or manipulation of the sample. Known since the Sixties, when the first lasers appeared, SP is the granular interference obtained when coherent light illuminates a rough object, with roughness of the same order of magnitude of the wavelength (around 400–700 nm). Traditionally, static SP has been exploited to study the roughness of surfaces [31], to measure the thermal strain and the elastic modulus of mechanical specimens [32,33], to detect the absolute position of objects [34], measure blood flow [35], and characterize the granularity of a powder surface [36]. Recently, SP imaging has grasped the interest of researchers for studying fluids. Indeed, also emulsions, suspensions, and opaque fluids in general can generate SP since they are constituted by particles floating in a liquid matrix. The particles contained in liquid suspensions are subject to Brownian motion; the SP they generate is dynamic and, thus, more challenging to be investigated. On this topic, for example, Héran et al. separately studied the p- and s-polarized SPs to measure the absorption and scattering coefficients of turbid fluids [37]. In [38], the authors analyzed polarized SP images produced by suspensions of polystyrene microspheres by comparing the results obtained with two different experimental deployments: a light transmission setup and a backscattering configuration. More recently, researchers have started exploiting artificial intelligence (AI) to extract information that images of dynamic SP are dense of [39,40]. For instance, Jakubczyk et al. applied convolutional neural networks (CNNs) to SP images generated from nanoparticle suspensions to classify SP images generated from 73 nanoparticle suspensions [41]. The classifier recognized nanoparticle material, size, and suspended phase concentration. In [42], Yan et al. used a transmission configuration to collect images of the SP produced by illuminating suspensions of plastic microspheres and milk powder with a He–Ne laser; then, they used a deep learning model for automatic recognition of samples with different particle concentrations, showing a good clustering capability. Endo et al. used a very similar configuration and applied CNNs to identify the size and concentration of microplastics [43]. However, these works only consider suspensions of plastic or silica particles with controlled geometry and dimension, prepared ad-hoc in the laboratory. In a previous work, we utilized SP imaging to test commercial milk and identify water-based milk adulteration [44]. The same group that authored paper [41], afterwards has also worked on the classification of off-the-shelf milk according to fat content [45] using the same CNN as in [41]. However, this work only focuses on commercial milk, pasteurized with ultra-high temperature (UHT) processes, which is far from exhibiting the same variability and complexity in composition as raw milk. Moreover, the authors did not provide additional information about the model uncertainty or about its explainability.

In contrast, our work is a thorough, yet preliminary, analysis of the applicability of SP imaging combined with AI to classify raw milk samples that are very similar in their nutritional contents, a scientific question that has never been addressed before. Our aim is to obtain a representation of the contents of the sample without quantifying the exact value of fats, proteins, and lactose. This is made possible by exploiting the SP imaging technique, since SPs contain information related to the sample composition. Even if the relationship between SP and milk composition is unknown, intelligent algorithms can help in recognizing different milk samples. By analyzing the SP images and extracting numerical parameters, we can describe the samples in a different feature space and find a way to link these features to the corresponding unique class (e.g., the sample label). A further and novel contribution is the metrological analysis of the method, allowing us to understand the reasons behind the classification outcomes. To this aim, we trained a feature-based model and an image-based model as examples to demonstrate the applicability of SP imaging and, at the same time, to explain why misclassification occurred in the models through uncertainty metrics. A simple and cost-effective optoelectronic configuration was adopted to generate and collect dynamic SPs of 20 raw cow milk samples, which are enough for our purposes. Classification of the samples was conducted on the basis of custom features used as predictors for an off-the-shelf Ensemble of Bagged Trees machine learning (ML) model, and compared with the performance of the CNN deep learning (DL) model described in [41].

2. Materials

2.1. Instrumental configuration

The configuration exploited to carry out experimental SP measurements is shown in Fig. 1. The coherent light source chosen to irradiate the samples is a semiconductor laser diode (L658P040, Thorlabs, NJ, USA) emitting red light (wavelength of 658 nm) at a maximum optical power of 40 mW. The diode is powered by a current

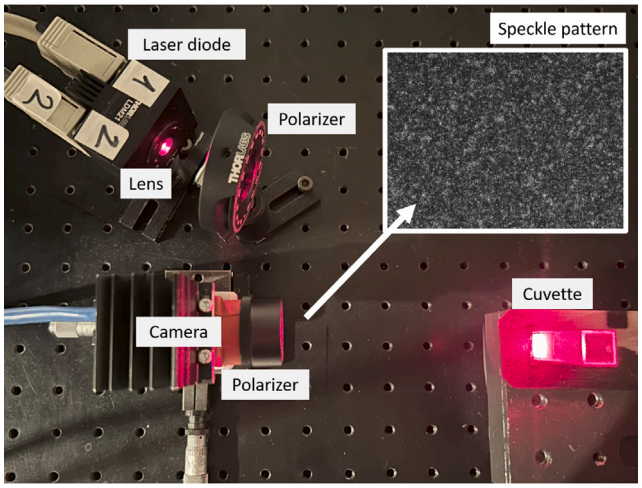

Fig. 1. Picture of the experimental setup with an example of an SP image obtained from a raw milk sample. Gray levels span from 0 (black) to 255 (white).

driver (LDC500, Thorlabs, NJ, USA) and is connected to a temperature controller (PRO800, Thorlabs, NJ, USA) for thermal stabilization. An aspheric lens with an anti-reflection coating (C230260P-B, Thorlabs, NJ, USA) is used to focus the light onto the external front side of the surface of a transparent polystyrene cuvette of volume 4.5 mL, and size $10\times10\times50$ mm, containing the milk sample, at an angle of about 30°. A linear polarizer (LPVISE100-A, Thorlabs, NJ, USA) is positioned before the lens to select only the main polarization component.

Images and videos of the SP generated by the scattering samples are acquired using a monochrome CMOS camera (Alvium 1800 U-158 m, Allied Vision, Germany) positioned in front of the cuvette. The camera orientation is selected to prevent the Snell reflection from the cuvette wall from reaching and saturating the CMOS sensor. The camera sensor has a size of 5.02×3.75 mm, a total number of pixels of 1456×1088, and a pixel size of 3.45×3.45 μm. A second linear polarizer is placed before the CMOS sensor. The camera is connected via USB to a laptop, allowing data acquisition through proprietary software (Vimba Viewer, Allied Vision, Germany). An example SP image obtained from a sample of raw milk is shown in the inset of Fig. 1.

2.2. Raw milk samples

In this work, we have analyzed raw milk, i.e. milk immediately after it is collected from the cows and not yet subject to any type of further processing. Raw milk is not pasteurized, homogenized, or sterilized, meaning that these types of samples deteriorate very quickly in an almost unpredictable way. As represented in Fig. 2, the scattering elements are constituted by the lipid globules (the dark yellow particles in the figure) and the caseins (the insoluble proteins of milk, represented by the brown elements in the figure), that float in the surrounding liquid phase constituted mainly by water, lactose, and soluble proteins. In raw milk, the dimension of fat globules is not uniform, since it is not homogenized, representing an additional factor of variability among the samples. It must be stressed that the dimension of the scattering elements also affects the resulting SP. Moreover, the higher the concentration of the lipid globules and caseins (hence, the higher the amount of scattering particles), the higher the amount of light back-diffused by the sample, thus leading to SP images with higher intensity. The strength of the light scattering also depends on the refractive index difference Δn between the particles and the surrounding liquid matrix. Indeed, while the refractive index n of the fat globules is around $n_{fat} \simeq$ 1.46–1.47 RIU and that of casein globules is around $n_{caseins} \simeq 1.57$ RIU [46], that of the liquid matrix n_{matrix} depends on the content of

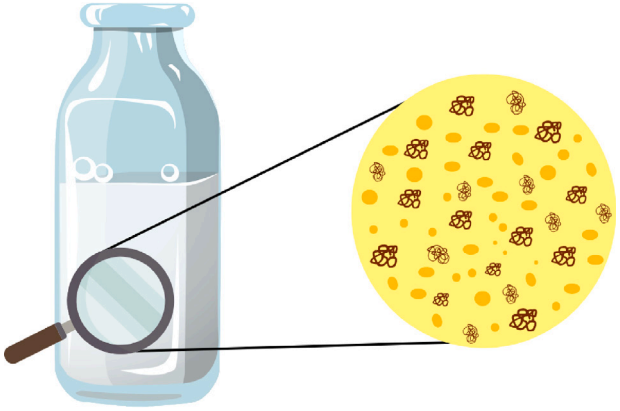

Fig. 2. Schematic representation of the scattering elements (lipid micelles and caseins) constituting raw milk samples.

water ($n_{water} = 1.33$ RIU) and on the concentration of lactose and soluble proteins. In more detail, the higher the content of the latter, the smaller the difference Δn, leading to a reduction of the back-scattered light according to Fresnel formulas and so to a reduction of SP intensity. Hence, the concentration of suspended elements (fat micelles and casein globules) and of soluble elements (lactose and soluble proteins) plays a counteracting effect on the generated SP: an increase of suspended element concentration determines more back-diffused light, while an increase of soluble element concentration tends to reduce the amount of back-scattered light. As a consequence, the average intensity, contrast, and spatial properties of SP images, described in detail in Section 3.1, are determined by the overall effect of this phenomena.

For our study, raw milk samples were collected from 20 different cows labeled with a numeric ID from 1 to 20 by the researchers of the *CERZOO - Centro di Ricerche per la Zootecnia e l'Ambiente Research Center* of Università Cattolica del Sacro Cuore (Piacenza, Italy). For each sample, fats, proteins, and lactose quantities were measured using a MilkoScan tester (Foss, Italy). The summary of these data can be found in Table 1. Proteins are divided into soluble and non-soluble.

2.3. Data collection procedure

A total of 4 acquisition Campaigns of experimental measurements were conducted on the day of the milking (D1), when all samples were collected, and the following day (D2). For all Campaigns, the laser current was set to 85 mA, resulting in an optical power of about 23 mW reaching the sample. The camera exposure time and frame rate were set to 800 μs and 50 fps, respectively. Such a value of the exposure time was chosen to reach a reasonable trade-off between sufficiently bright and clear SPs and not-saturated images. It must be noted that the chosen value of the exposure time is much longer than the typical SP decorrelation time for our specific experimental condition (of the order hundreds or even tenths of nanoseconds, due to multiple scattering and highly concentrated raw milk). Hence, with our low-cost industrial camera we can collect only uncorrelated partially developed [47] SP images, resulting from the sum of fully developed SPs, integrated over the exposure time [47]. However, this does not represent an issue since our entire analysis is based on the investigation of the statistical and spatial properties of the SP images rather than the temporal evolution of the SP.

The 20 raw milk samples were collected from the stable on day D1 and taken to the laboratory for testing right after milking. The 20 samples were always stored in a refrigerator at $T = 4$ °C, as suggested in [48]. For each acquisition Campaign, a few milliliters of liquid were taken out to fill 3 rectangular cuvettes per milk sample, as shown in Fig. 3. Data acquisition Campaigns were designed and carried out as follows:

Table 1
Nutrient composition of the samples. Fats, non soluble proteins (N-S.) and soluble proteins (S.), and lactose are expressed as (g/L) and were measured with a MilkoScan.

Sample ID	Fats	N-S. Proteins	S. Proteins	Lactose
Sample_01	51.2	37.1	9.3	43.4
Sample_02	49.2	27.9	7.0	48.9
Sample_03	48.8	36.5	9.1	48.5
Sample_04	46.2	29.1	7.3	49.1
Sample_05	45.2	27.8	7.0	48.9
Sample_06	45.0	29.3	7.3	48.8
Sample_07	44.4	27.5	6.9	48.7
Sample_08	41.2	29.5	7.4	48.0
Sample_09	40.5	30.8	7.7	49.0
Sample_10	40.4	30.8	7.7	48.9
Sample_11	38.8	27.7	6.9	45.6
Sample_12	38.5	29.9	7.5	50.4
Sample_13	37.8	28.2	7.0	50.3
Sample_14	37.5	26.4	6.6	46.4
Sample_15	33.3	29.1	7.3	48.8
Sample_16	32.8	24.3	6.1	51.1
Sample_17	32.5	27.2	6.8	49.9
Sample_18	31.7	24.8	6.2	48.7
Sample_19	30.1	28.6	6.7	50.9
Sample_20	26.4	28.5	7.1	49.4

1. **Campaign 1:** temperature of the cuvettes was set to $T_1 = 22$ °C. The measurements were conducted on the morning of D1.
2. **Campaign 2:** temperature of the cuvettes was set to $T_2 = 22$ °C. The measurements were conducted in the afternoon of D1.
3. **Campaign 3:** temperature of the cuvettes was set to $T_3 = 10$ °C. The measurements were conducted in the afternoon of D1.
4. **Campaign 4:** temperature of the cuvettes was set to $T_4 = 22$ °C. The measurements were conducted in the morning of D2.

It is worth noting that only the liquid inside the cuvettes was taken outside the refrigerator. For Campaigns 1, 2, and 4, the target temperature of 22 °C was reached by leaving the cuvettes at room temperature. For Campaign 3, the cuvettes were stored in a refrigerated environment until the liquid temperature reached 10 °C. The samples temperature was measured before each acquisition using a thermometer. These two working temperatures were chosen because cows are usually milked in stables where the average temperature, in different seasons, typically varies in the range 7–25 °C [49]. In addition, by testing samples at $T_3 = 10$ °C the effect of low temperature on the measurements was studied in comparison. Indeed, it is well known from the scientific literature that the temperature affects the Brownian motion of the scattering particles (lipidic micelles and caseins) in the suspension, and, hence, the resulting SP images [50].

A brand-new cuvette was used each time a new acquisition was conducted (so, for each sample, during each Campaign), for a total of 4 Campaigns × 20 raw milk samples cuvettes. The sample preparation was always conducted by gently shaking the raw milk sample to prevent particle precipitation, but without leading to foam formation. Acquisitions were repeated 3 times for each milk sample, collecting 100 frames each time (every group of 100 SP images will be called "set" from now on for simplicity), for a total of 6000 frames per Campaign (100 frames × 3 sets × 20 milk samples). This results in a dataset of 24,000 frames (6000 frames per Campaign × 4 Campaigns).

3. Methods

Image analysis, feature extraction, and ML and DL model training and testing were conducted on a laptop equipped with an Intel i5 CPU of 1.60 GHz, 8 GB of RAM, 500 GB of HDD, and Windows 10 Pro. Moreover, a detailed analysis of the models' performance is presented. The software used was MATLAB 2024b (Mathworks Inc., Natick, MA, USA).

3.1. Features extraction

SP are monochromatic images that do not directly represent or visualize the internal fluid structure or composition. However, the granular pattern distribution of the SP images and their statistical properties are strongly related to the sample content in terms of lipids, proteins, and carbohydrates, as previously described in Section 2.2. It is also known from the literature that concentration, dimension and structure of fat globules, proteins, lactose molecules and micelles formed by aggregation of caseins with calcium phosphate simultaneously contribute to the determination of the properties of scattered light [45,51–53]. Hence, SP images were pre-processed to extract relevant statistical and image-based features that were then used as input for the ML models. A total of 16 features were extracted; 7 of them were selected according to the existing literature on statistical features for SP analysis [39,54], 4 extracted from the Gray-Level Co-Occurrence Matrix (GLCM) computed from each SP image, and 5 related to the distribution of the bright speckle grains and dark regions in the images, obtained after the application of image processing techniques. These features were chosen since they demonstrated promising results in a previous work [55].

The first 7 statistical features used are: (i) the image mean intensity (computed as the mean gray level value in the range [0 – 255]), (ii) the standard deviation of the image intensity, (iii) the image median, (iv) the image kurtosis, (v) the image skewness, and (vi–vii) the dimension of the SP grains along the x- and y-direction. These two parameters can be retrieved as the Full Width at Half Maximum (FWHM) of the autocovariance function of the SP image, which results in two values referring to the vertical and horizontal size of the SP grains (FWHMx and FWHMy, respectively). The autocovariance function, obtained from the calculation of the autocorrelation function, is commonly used in SP analysis [47,56].

The GLCM of an image represents the distribution of co-occurring pixel values at a given offset. It computes how often pairs of pixels with a specific intensity value and offset occur in the image, capturing patterns within it [57]. The $(i, j)_{th}$ element of the GLCM represents the number of times a pixel with intensity i is neighbor to a pixel with intensity j. The GLCM is often used to analyze the texture of images and is a powerful tool in SP analysis. A GLCM was computed from every SP frame collected and then, from the GLCM sparse matrices, the following parameters were extracted: (i) GLCM contrast, (ii) GLCM correlation, (iii) GLCM energy, and (iv) GLCM homogeneity. GLCMs and their properties were computed using the MATLAB functions *gracomatrix* and *graycoprops*, based on the formulations in [58].

The last 5 features were retrieved by applying an image processing analysis aimed at extracting features related to the inherent shape, dimension, and distribution of the bright and dark areas of the SP by means of blob analysis, which is a fundamental computer vision technique used to analyze the contents of an image [59]. First, the original SP image was binarized to enhance black and white regions. Binarization was performed using MATLAB's *imbinarize* function, which automatically determines the binarization threshold to a value that minimizes the intraclass variance between thresholded black and white pixels [60]. Second, a blob analysis technique was applied to the binarized image to detect regions called "blobs" (i.e. regions of connected pixels with the highest gray level value), which share common properties. It was observed that SP images contain wide regions with curved non-regular shapes and small regions almost resembling circles. The features extracted thanks to the blob analysis are: (i) number of blobs, (ii) mean area of the blobs in pixels, (iii) mean circularity of the blobs, (iv) mean eccentricity of the blobs, and (v) mean orientation of the blobs.

To reduce potential discrepancies among Campaigns and features with very different ranges and absolute values, all features were standardized using the z-score method [61].

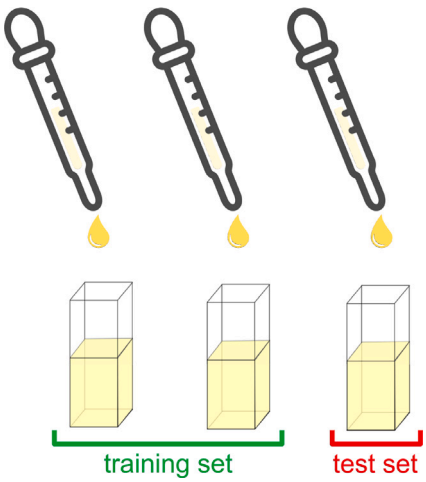

Fig. 3. Schematic representation of the cuvettes prepared for each sample during each Campaign. Data belonging to two out of three were included in the training datasets, while the third one was included in the test datasets. The cuvettes were randomly selected.

3.2. Datasets creation and training–testing strategy

The features explained in Section 3.1 and the original SP images were organized in training and testing datasets as summarized in Table 2. To divide the datasets into training and testing, for each sample of each Campaign, the algorithm randomly selected one set among the available three (each set contains 100 SP, as described in Section 2.3). The selected set was included in the testing dataset, while the remaining two were included in the training dataset (see Fig. 3). This ensured a balanced composition of the two datasets (about 66% of data used for training and 33% used for testing).

It is worth noting that the datasets used for ML analyses contain the 16 statistical features only, while the datasets used for DL analysis contain the original SP images center-cropped to size 640 × 480. Due to the properties of SP, the information contained is preserved even after cropping, thus reducing the training time of the DL models without losing information. Images were obtained by illuminating the sample with a red laser, but they are grayscale images (since they were acquired with a black-and-white CMOS camera), suitable for DL training. Keeping this in mind, for the sake of simplicity, we will refer to the dataset names in the remainder of this paper without specifying what data are contained in them.

All the different datasets that we have considered and investigated in our work are summarized in Table 2. Datasets containing the standardized features are represented by subscripts "α", "β", and "$\beta 2$", while those containing standardized SP images are indicated by subscripts "γ", "δ", and "$\delta 2$", respectively, as specified in Table 2. Data from Campaign 3 (collected at sample temperature $T_3 = 10$ °C) were excluded from datasets "β" and "δ" but included in datasets "α" and "γ", allowing us to test the impact of the sample temperature on prediction accuracy. Data standardization of all datasets was conducted using z-score: data contained in the training datasets were standardized (every feature on its own) using the mean and standard deviation obtained from the training datasets themselves. On the other hand, data in the test datasets were standardized using the same mean and standard deviation, computed on the corresponding training datasets as indicated in Table 2. It is worth noting that test datasets "$\beta 2$" and "$\delta 2$" include data from all Campaigns (as test datasets "α" and "γ", respectively), but were standardized using the mean and standard deviation of training datasets "β" and "δ", respectively. This approach allowed us to study what happens if a ML/DL model is used to test data collected at a different temperature, data that the models have never had the opportunity to learn during the training step. training–testing strategies are summarized in Table 3, where "ML" stands for Machine Learning models (structure detailed in Section 3.3.1), and "DL" stands for Deep Learning models (structure detailed in Section 3.3.2). More precisely, ML1 was obtained by training the model on dataset Training_α and testing it on Testing_α, ML2 by training on Training_β and testing it on Testing_β, and ML3 by training on Training_β and testing on $\text{Testing}_{\beta 2}$. Similarly, DL1 was obtained by training on dataset Training_γ and testing it on Testing_γ, DL2 by training on Training_δ and testing it on Testing_δ, and DL3 by training on Training_δ and testing it on $\text{Testing}_{\delta 2}$.

Table 2
Summary of the composition of the datasets. Instances refer to the number of data elements in the corresponding datasets. Both features and speckle pattern images were standardized using z-score.

Dataset	Campaigns	Data type	Instances	Z-score on?
$Training_\alpha$	1, 2, 3, 4	16 Features	16,000	$Training_\alpha$
$Training_\beta$	1, 2, 4	16 Features	12,000	$Training_\beta$
$Training_\gamma$	1, 2, 3, 4	SP images	16,000	$Training_\gamma$
$Training_\delta$	1, 2, 4	SP images	12,000	$Training_\delta$
$Testing_\alpha$	1, 2, 3, 4	16 Features	8000	$Training_\alpha$
$Testing_\beta$	1, 2, 4	16 Features	6000	$Training_\beta$
$Testing_{\beta 2}$	1, 2, 3, 4	16 Features	8000	$Training_\beta$
$Testing_\gamma$	1, 2, 3, 4	SP images	8000	$Training_\gamma$
$Testing_\delta$	1, 2, 4	SP images	6000	$Training_\delta$
$Testing_{\delta 2}$	1, 2, 3, 4	SP images	8000	$Training_\delta$

3.3. Classification models

3.3.1. Machine learning models

MATLAB Classification Learner was used to conduct this analysis. The model of choice was an Ensemble of 30 Bagged Decision Trees. Bagged decision trees are a popular type of ML model because the problem of high variance affecting traditional decision trees is tackled by bootstrap aggregation (or bagging), a general-purpose procedure for reducing the variance of a statistical learning method. Ensembling is a methodology to improve the overall performance of a statistical model by training multiple learners (in our case, 30) and then combining them in a single model [62]. Decision trees are quick and easy ML models that are easily interpretable, avoiding the "black box" structure typical of neural networks. This is the main reason why the research community working on milk adulteration often adopts them [63,64]. Furthermore, for the goal of understanding classification uncertainty, decision trees are the best choice as highlighted in [65].

The model was automatically validated using cross-fold validation with 5 folds. This architecture was used for the ML training–testing combinations presented in Section 3.2 resulting in models ML1–ML3.

3.3.2. Deep learning model

Among the available DL models trained to classify image contents (which are based on the convolution operation and are known as Convolutional Neural Networks or CNNs), the network described in [41] was specifically designed to work on SP images. These authors used it to distinguish colloidal suspensions of nanoparticles diluted in distilled water with remarkable results. For the purpose of this work, the original network described in [41] was modified and re-trained from scratch using the training datasets "γ" and "δ". The input images were of size 640 × 480 pixels, center-cropped from the original full-resolution SP images, as described in Section 3.2. Pixel intensities contained in the SPs were standardized using z-score according to the model (e.g. standardization mean and standard deviation obtained from the training dataset were applied on the corresponding test dataset, as reported in Table 2). Pre-processed images were saved in .tif format to keep the negative values resulting from the standardization process. The network has 35 layers with 8 convolutional layers (containing 64 3 × 3 filters each), 4 max-pooling layers with 3 × 3 filter size, 7 batch

normalization layers, and 2 fully connected layers composed of 512 and 256 neurons, respectively (output size equal to the number of classes, corresponding to the 20 samples ID). Training hyperparameters were: mini-batch size of 16, maximum epochs 30 with 1000 iterations each, learn rate schedule piecewise, initial learning rate 0.01, shuffle every epoch, learning rate drop period 10, optimization algorithm Stochastic Gradient Descent with Momentum set at 0.9. This model architecture was used for the DL training–testing combinations presented in Section 3.2, resulting in models DL1–DL3.

3.4. Uncertainty analysis

The analysis and the quantification of the prediction process are fundamental to assessing the level of trustworthiness of the models. Several factors contribute to uncertainty in classification tasks, such as measurement noise, observation conditions, and the choice of the classification model. However, in classification, the main issue is the determination of class separability. Generally, an object is considered part of a certain class C if described by features common to that class. Hence, the more separable the classes according to these features, the easier it is for a model to distinguish between them with relatively low uncertainty. When two classes have similar features, data points belonging to them become ambiguous because they can belong to both classes from the model point of view, thus increasing its prediction uncertainty. This is especially true in the case of data points belonging to datasets unseen by the model during training. Considering a certain SP image obtained from a milk sample, the sample can be represented in a high-dimensional space either by the custom 16 features described in Section 3.1 or by the pixel intensities of the SP images. If another sample has a similar composition, we expect that the resulting SP image will be similar as well, and thus its representation in the high-dimensional space. Through uncertainty analysis, we can understand if the model can recognize the two samples as different or not, and, eventually, we can provide a measure of doubt related to the prediction. In this context, the work of Chlaily et al. [66] was a pivotal step in defining metrics that can answer two fundamental questions related to classification models: (i) how far is the model from a random guess, answered by geometry-based uncertainty metrics, and (ii) the type of confusion, answered by homophily-based uncertainty metrics.

A general classification model is trained to classify the input data x into one of the C classes defined in the model structure. After analyzing the input x, the model outputs a probability vector P_x of size $1 \times C$, in which its ith elements P_i contain the probability of the input x of being classified as belonging to class C_i. The sum of all the i elements in P_x is equal to 1. When the classifier makes a prediction, it assigns to x its predicted class by selecting the maximum of P_x. Ideally, P_x contains only one class with probability equal to 1 while the others will be 0. This is the case when the classifier is perfectly certain of its decision outcome. The most uncertain case is when the probability values in P_x are all the same, equal to $1/C$. In this case, the classifier cannot make a prediction, so it makes a random guess. It is important to stress that the uncertainty of the classifier is unrelated to its classification accuracy, meaning that the classifier can be certain of the prediction, but the predicted class can be completely wrong, or the prediction can be correct even if the classifier is uncertain about it. Please note that classification accuracy is generally computed as the number of correct classifications over the total number of predictions, and it is not the same as measurement accuracy.

Geometry-based metrics, as defined in [66], consider the prediction vector P_x as an item in the space of all the possible probability combinations, where the extremes correspond to certain cases where only one among the C classes has a probability equal to 1, and the uncertain case corresponds to the case when all probabilities are equal to $1/C$. Hence, geometry-based metrics aim to compute the distance of P_x from the uncertain case (random guess).

Homophily-based metrics, on the other hand, try to explain why the model makes wrong predictions. By considering the original distribution of features that describe the inputs, mistaking two classes with similar distributions is acceptable, while mistaking classes with no common characteristics suggests a wrong model choice or training strategy.

In this work, the geometry-based uncertainty metrics (GU) adopted are (i) the Normalized Gini–Simpson index [67] GU_{GS}, and (ii) the normalized Shannon entropy [68] GU_{SE}. Both are particular cases of Euclidean and Kullback–Leibler uncertainty metrics, respectively, as defined in [66]. The homophily-based uncertainty metric HU is computed according to the definition of the Energy distance [69] as in [66]. For the general mathematical definitions of metrics and their proof, the reader is encouraged to refer to the original article [66]. The software used to conduct uncertainty analysis was a modified version of the original code published on GitHub by authors of [66], and it was developed in Python 3.8 on the same machine described in Section 3.

3.4.1. Geometry-based metrics

After obtaining the trained models ML1–ML3 and DL1–DL3, each instance x of the corresponding test dataset is elaborated to produce P_x of size $1 \times C$, where $C = 20$. P_x contains the per-class probability P_i of the input x of being classified as belonging to class C_i. According to [66], the geometry-based uncertainty metrics used in this work are computed as follows, for a given input x:

$$GU_{GS} = 1 - \frac{\sum_{i=1}^{C} (P_i - \frac{1}{C})^2}{(1 - \frac{1}{C})} \tag{1}$$

$$GU_{SE} = 1 - \frac{\sum_{i=1}^{C} P_i \cdot \log(C \cdot P_i)}{\log(C)} \tag{2}$$

These formulas are applied considering the predictions obtained by each trained model ML1–ML3 and DL1–DL3 on their specific test dataset individually. The results are grouped per ground-truth class, thus obtaining the average values of GU_{GS} and GU_{SE} with the corresponding standard deviation.

3.4.2. Homophily-based metrics

The type of confusion depends on the similarity of the class distributions (e.g. their probability density function, or PDF) in the feature space and reflects how likely it is to confuse them. In our specific case, for ML models the data's PDF corresponds to the standardized features described in Section 3.1, while for DL models the data's PDF corresponds to the pixel intensities stored in the images and standardized according to the procedure described in Section 3.3.2.

There can be two types of misclassification: (1) the model confuses two classes that are similar (e.g., close in the feature space), and (2) the model confuses two classes that are different (e.g., distant in the feature space). To these two cases, a different weight (or penalty) should be assigned. For case (1), the PDF of confused classes is similar, reflecting that the phenomenon represented by them is hard to distinguish, so the assigned weight should be low. On the other hand, for case (2), the model confuses classes with different PDFs, highlighting that the model chosen is not the best for the phenomenon object of the study, or that it was not trained correctly. Hence, the assigned weight should be high. These weights are represented by the pairwise distance between C classes (with C ranging from 1 to the maximum number of classes the model is trained to recognize), quantified by the Energy distance [69], which is a similarity measure between the classes' PDF. Energy distance between classes is calculated on the corresponding test dataset and represented in the form of an Energy matrix $\mathbf{H}$, in which values close to 0 represent classes that have similar PDFs, while values close to 1 represent the opposite case. By definition, $\mathbf{H}$ is a symmetric matrix with non-negative elements and zero values on the main diagonal. The computation of $\mathbf{H}$ is conducted considering the classes' PDF; hence, in our case, the 16 features for ML models, and the pixel intensities of

the original SP images for DL models. To effectively use the original SP images as PDFs for DL models, they are treated as one-dimensional vectors thanks to a rolling operation (e.g., all the rows of the image matrix are procedurally moved one next to the other, forming a vector of 1 row and N_{px} columns, where N_{px} represents the total number of pixels in the image). Considering that the frames used for the training of DL models were center-cropped to a resolution of 640 × 480 pixels, N_{px} is equal to 307,200 pixels.

The procedure to calculate the energy matrix **H** was as follows:

1. *Input data creation:* A matrix **X** containing the features of all the instances in the test dataset of choice is created. The matrix has size $N_{instances} \times N_{features}$, where $N_{instances}$ corresponds to the number of instances in the test dataset (see Table 2) and $N_{features}$ is either equal to 16 in the case of ML models or to the total number of pixels in the SP, N_{px}, for DL models. Then, the ground-truth vector **Y** is created, containing the true class number from 1 to 20 of the milk samples. This vector has size $N_{instances} \times 1$. Please note that matrix **X** and vector **Y** have the same order of rows; thus, the features in **X** at row i correspond to the ground-truth in **Y** at row i.
2. *Selection of pairs:* The algorithm creates two sub-matrices, S_i and S_j, by selecting the rows of **X** that correspond to class i and j, respectively. To check if a specific row of **X** belongs to a certain class, the data in **Y** is used. As explained in Section 3.2, the test datasets contain 100 instances for every specific class and for each Campaign, corresponding to 400 instances for datasets "α", "$\beta 2$", "γ", and "$\delta 2$"; and 300 instances for datasets "β" and "δ". By calling this value N_{test}, it is straightforward to obtain the size of S_i and S_j as $N_{test} \times N_{features}$.
3. *Pairwise Energy computation:* The Energy distance is computed using the SciPy function "energy_distance" belonging to the statistical package [70], considering iteratively all the elements of S_i and S_j corresponding to a certain feature in column f (from 1 to $N_{features}$). This operation produces a vector $\mathbf{D_H}$ of size $1 \times N_{features}$. The value of **H** corresponding to the current pairwise computation (i, j) is obtained as $\mu(\mathbf{D_H}) + \sigma(\mathbf{D_H})$, where $\mu(\mathbf{D_H})$ correspond to the mean of vector $\mathbf{D_H}$, and $\sigma(\mathbf{D_H})$ to its standard deviation.

Then, according to the definition in [66], HU is computed as:

$$HU = \frac{\mathbf{P}^T \cdot (\mathbf{H} \odot \mathbf{H}) \cdot \mathbf{P}}{\mathbf{P}_{max}^T \cdot (\mathbf{H} \odot \mathbf{H}) \cdot \mathbf{P}_{max}} \tag{3}$$

where **P** corresponds to the matrix of all the predictions P_x and it is of size $N_{instances} \times 20$, where $N_{instances}$ is the size of the test dataset considered. The symbol $\odot$ denotes the Hadamard product and superscript T the transpose operation. $\mathbf{P}_{max}^T$ is obtained as $argmax(\mathbf{P}^T \cdot (\mathbf{H} \odot \mathbf{H}) \cdot \mathbf{P})$. Please note that the abovementioned formula is applied considering the predictions obtained by each trained model ML1–ML3 and DL1–DL3 on their specific test dataset individually. The results are grouped per ground-truth class, thus obtaining the average values of HU with the corresponding standard deviation.

4. Results and discussion

As a preliminary step toward milk identification, a Principal Component Analysis (PCA) [71] was conducted on the standardized features of each Campaign individually to visualize the distribution of the milk sample data according to the first two principal components (PC). The PCs are a new representation of the original reference system, allowing the data to be described in a lower-dimensional space (e.g. from space 16 to space 2) without losing information. They are calculated by MATLAB and ordered according to the percentage of data variance they explained; hence, by choosing the first two, we selected the PC that explained more than 90% of data variance. The result of this analysis is shown in Fig. 4 for Campaign 2 (T_2 = 22 °C) and Campaign 3 (T_3 = 10 °C). The PCA results of the other two Campaigns were omitted for brevity since they are similar.

Table 3
Summary of the validation and testing results for models ML1–ML3 and DL1–DL3. Model performance is assessed using prediction accuracy on the test dataset of each model [%]. Processing and training time of all models is shown for quick reference.

Training	Test	Model	Accuracy	Proc./Training time
$Training_{\alpha}$	$Testing_{\alpha}$	ML1	95%	4 h / 20 min
$Training_{\beta}$	$Testing_{\beta}$	ML2	94%	4 h / 20 min
$Training_{\beta}$	$Testing_{\beta 2}$	ML3	71%	4 h / 20 min
$Training_{\gamma}$	$Testing_{\gamma}$	DL1	95%	2 h / 30 h
$Training_{\delta}$	$Testing_{\delta}$	DL2	94%	2 h / 30 h
$Training_{\delta}$	$Testing_{\delta 2}$	DL3	72%	2 h / 30 h

The PCA analysis of Campaigns 2 (Fig. 4(a)) shows distinguishable round clusters, but several outliers belonging to different classes appear overlapped. Clusters in Campaign 3 (Fig. 4(b)) are more separated, showing an elliptical shape, and appear less spread, thus with a lower inter-class variance. This behavior is likely due to the lower temperature of the samples during the measurement (T_3 = 10 °C), leading to a slightly less relevant impact of Brownian motion. Performing the same data analysis using the first three PCs resulted in similar outcomes, with previously overlapping clusters slightly more separable thanks to the third coordinate. These outcomes suggest that simple clustering algorithms could separate some classes from the others, especially if applied to data of individual Campaigns, but they are not sufficiently powerful to robustly distinguish each class. This is even more evident considering the differences between the clustering obtained for each Campaign, mainly due to slightly different environmental conditions. This reflects the intrinsic variability of the measurement when conducted on raw milk, which in turn affects the formation of SP images. In addition, the similarity between clusters of milk samples depends mostly on their nutritional contents, but also on a combination of several factors, including lactation stage, days of lactation, cow's weight, and the fact that they are fed the same way and live on the same farm, not to mention physical properties of raw milk affecting the formation of SP images (for example, the sample temperature). Since the desired inspection method should be able to differentiate samples regardless of the acquisition Campaign, ML and DL models are confirmed to be the most suitable analysis techniques for the task.

4.1. Classification results

All models were trained according to the training–testing strategy presented in Section 3.2 nad 3. Prediction accuracy was used to assess the model performance, which is defined as the number of correct predictions of a given class over the number of ground-truth occurrences of the class in the dataset. Results are summarized in Table 3 for both ML and DL models, including processing and training times required for each model for quick reference. ML1, ML2, DL1, and DL3 led to a very high overall accuracy of 95% (ML1–DL1) and 94% (ML2–DL2), respectively. On the other hand, for ML3 and DL3 (for which Campaign 3 was excluded from the training but included in the test), prediction accuracy notably drops on the corresponding test dataset to 71% and 72%, respectively. This result was expected and confirms that data collected at different sample temperatures cannot be correctly classified if they are not included in the training of the models.

Confusion matrices of the best-performing models (ML1 and DL1) are shown in Figs. 5(a) and 5(b), and of the worst-performing models (ML3 and DL3) are shown in Figs. 5(c) and 5(d). For all models and for all classes, four extra prediction metrics are summarized in Fig. 6 for ML and DL models, respectively. The prediction metrics considered are: the true positive rate (TPR), also known as sensitivity, the false negative

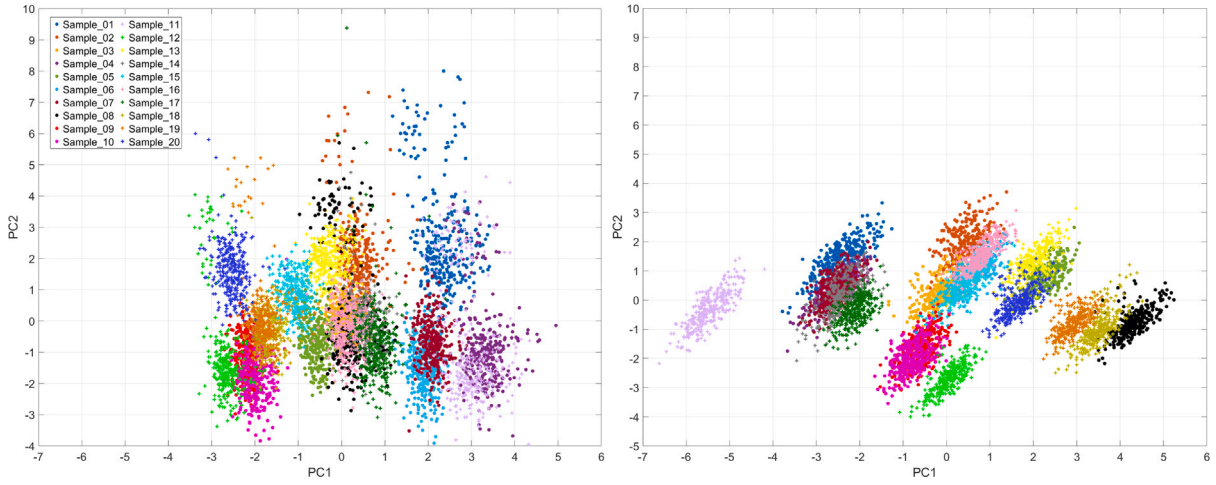

Fig. 4. Result of the PCA analysis applied to all the features of the cow milk samples for each Campaign individually. (a) Campaign 2 ($T_2 = 22$ °C), (b) Campaign 3 ($T_3 = 10$ °C).

rate (FNR), the positive predicted value (PPV), and the false detection rate (FDR). TPR is the probability that an actual positive will test positive, while FNR represents the probability that a true positive will be missed by the test. PPV and FDR offer further insights into the results: PPV represents the probability of the prediction being correct among all true positive and true negative results, while FDR represents the number of false positive classifications to the total number of positive classifications. It is worth mentioning that $TPR+FNR = PPV+FDR = 100\%$. These metrics can be derived from the confusion matrix as the row (TPR and FNR) and column (PPV and FDR) summaries, normalized over the corresponding row or column.

Interestingly, for both ML3 and DL3 (Figs. 5(c) and 5(d)) the confusion matrices show that several milk samples are classified as if they were Sample_12, but, on the other hand, correctly identifies the occurrences of Sample_12 in almost 100% of cases. Moreover, in ML3, Sample_09 and Sample_12 have quite high FDR (51% and 78%, respectively), indicating that the model often predicts those two classes when the ground truth was a different class. The same happens for DL3 in the case of Sample_09 (34%), Sample_10 (34%), Sample_12 (70%), Sample_18 (60%), and Sample_19 (50%). By looking at the clusters in Fig. 4, Sample_12 does not appear to overlap much with other clusters, while the other four are overlapped in couples (Sample_09 overlaps with Sample_10, Sample_18 overlaps with Sample_19), especially in Campaign 2 (Fig. 4(a)). This suggests that the composition of these samples produces similar SP features; in fact, samples with better classification accuracy overlap only in some Campaigns (e.g., Sample_04 and Sample_14 overlapping in Campaign 2 but not in Campaign 3, as shown in Fig. 4).

Overall, the performance of the two architectures is almost identical, so the choice of the right approach should be driven by other factors, such as the simplicity of training, the computational requirements needed, or the model's uncertainty. DL models required a conceptually easier pre-processing procedure (e.g., cropping of the central region of the frames and SP z-score standardization), but stricter hardware requirements (e.g. disk space to store the standardized SP, suitable processing units to run the DL model). Overall, the pre-processing step needed by DL models took 10 h per dataset and 55 Gb of disk space, and the model training took approximately 30 h per model. On the other hand, ML models require the user to conduct a thorough feature engineering process, corresponding to a processing time of around 4 h per model to compute the standardized features, but the produced feature tables took only 20 Mb of disk space and the model training required only 20 min each. Moreover, the inference time (i.e., the time required by a trained model to compute a single prediction on a new sample) of DL models is around 100 ms, in contrast with 40 ms for ML models.

4.2. Results of uncertainty analysis on trained models

As extensively explained, some milk samples have very similar composition (see Table 1). This translates in SP similar characteristics (either features for ML or pixel intensities for DL) and, in turn, affects the model performance. This is precisely why studying the uncertainty of the model is important. First, Energy matrices were computed for all test datasets: "α", "β, and "$\beta 2$" for ML models ML1, ML", ML3, "γ", "δ", and "$\delta 2$" for DL models DL1, DL2, DL3. For sake of brevity, only Energy matrices related to the test datasets "α" ($\mathbf{H}_{\alpha}$) and "γ" ($\mathbf{H}_{\gamma}$) related to datasets "β", "$\beta 2$", "δ", and "$\delta 2$" show similar values. The two matrices containing normalized values in range $[0, 1]$ are shown as heat maps in Figs. 7(a) and 7(b), respectively. In the case of $\mathbf{H}_{\alpha}$, it is highlighted that Sample_18, Sample_19, and Sample_20 are similar because they have a low Energy distance, just as Sample_09, Sample_10, and Sample_12. Other classes present quite low Energy values as well, for example, Sample_14, Sample_16, and Sample_17. However, for $\mathbf{H}_{\gamma}$ the same considerations are not true since it appears mostly green (values greater than 0.5). It is also interesting to note that those pairs that exhibited high Energy (pairs with no similarities) in $\mathbf{H}_{\alpha}$ do the same in $\mathbf{H}_{\gamma}$ (e.g., Sample_01, Sample_04, Sample_07, and Sample_11 are not similar to Sample_18, Sample_19, and Sample_20). In addition, by checking the rows of $\mathbf{H}_{\alpha}$ and $\mathbf{H}_{\gamma}$ it is possible to visualize which class is generally similar to the others. For example, by looking at $\mathbf{H}_{\alpha}$, even if Sample_18, Sample_19, and Sample_20 are similar, they appear to be distant from the other classes. On the other hand, Sample_05 and Sample_03 often present values around 0.5 or lower, suggesting that they share similarities with a lot of classes. However, this is not the case for $\mathbf{H}_{\gamma}$, for which it is evident that similar classes are just a few and most of them have an Energy value around 0.5. This results in higher penalties in the computation of uncertainty, as described in Section 3.4.

Results of the three uncertainty metrics (GU_{GS}, GU_{SE} and HU) are shown in Fig. 8 for ML models and in Fig. 9 for DL models, where different colors indicate different models. The average value per class is shown as a colored bar, and the standard deviation is the error bar (only the positive is shown since negative probability is meaningless). A value of 100% indicates maximum uncertainty. We recall that the uncertainty metrics described in Section 3.4 are obtained by considering, for each trained model ML1–ML3 and DL1–DL3, the predictions computed on

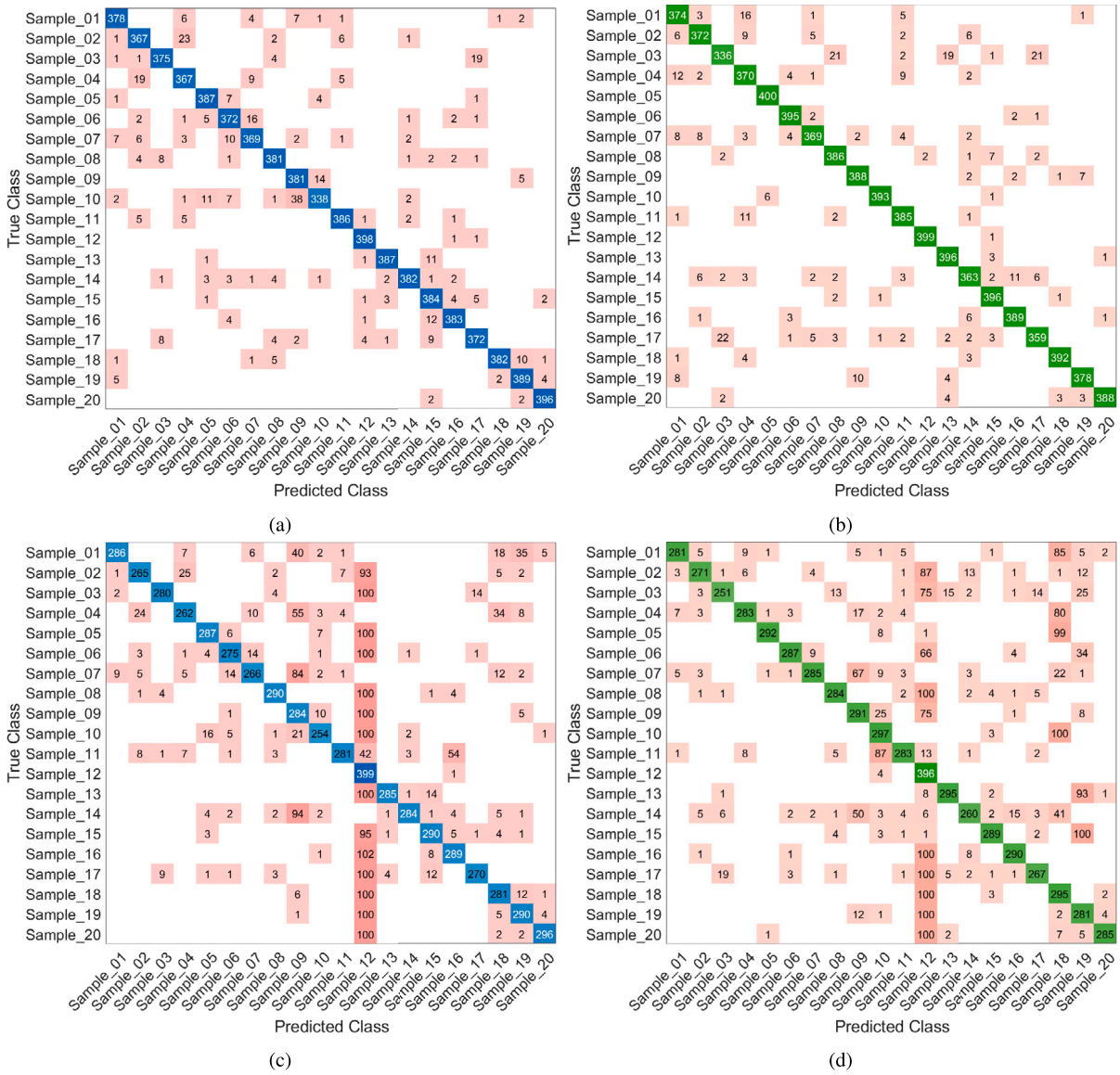

Fig. 5. Confusion matrices of the model performance achieved on the test set for the best-performing models, (a) ML1 and (b) DL1, and the worst-performing models, (c) ML3 and (d) DL3.

the corresponding test dataset according to Table 2. Values close to 100% correspond to higher uncertainty.

Since geometry-based uncertainty metrics GU_{GS} and GU_{SE} represent the distance of the model from a random guess (case of maximum uncertainty with all elements of the prediction vector P_x equal to $1/C$), low uncertainty means that the model is often sure of its predictions, regardless of their correctness. GU_{GS} is on average below 20% for ML1 and ML2 with almost double standard deviation. For ML3 results are worse, in a few cases reaching almost 30% (doubled if considering the standard deviation). This was expected since ML3 is tested on dataset "$\beta 2$" which includes data from Campaign 3 that was excluded from the training. For GU_{SE}, the values are generally halved compared to GU_{GS}, typically under 20% if considering the standard deviation. Notably, for DL models GU_{GS} and GU_{SE} are generally comparable with the ML counterparts, reaching lower values overall (except for Sample_14 for which the values are worse in both cases). This suggests that the DL models are more confident of their predictions compared to ML models. Interestingly, the performance of the Sample_12 class is always the best for all models, reflecting the behavior already observed in the confusion matrices and prediction metrics. It is interesting to note that geometry-based metrics are related to FNR values reported in Fig. 6. Generally, classes for which FNR is lower than 6% also show notably lower uncertainty values, as in the case of Sample_12 and Sample_20 for all ML models, and for Sample_05 and Sample_12 for all DL models.

As for the homophily-based uncertainty HU, it is interesting to note that it follows the general trend of geometric-based uncertainty for both ML and DL models: classes that showed high GU_{GS} and GU_{SE} values also show high HU values and vice versa. According to [66] and to Eq. (3), high HU values indicate that the model confuses distant classes. This is also explained by how the Energy matrix is used in the formula, which assigns a higher penalty the more distant the classes are (corresponding to an Energy value close 1). Considering the matrices in Fig. 5, and the Energy matrices in Fig. 7, HU values in Fig. 8(c) and Fig. 9(c) are mostly due to confusions between close classes (red color in $\mathbf{H}_{\alpha}$ and $\mathbf{H}_{\gamma}$) than distant ones, which are often classified correctly. This reflects the fact that the many samples have similar nutritional contents (as highlighted by the Energy matrices in Fig. 7.) However, when misclassifications between distant classes happen, their contribution toward the total uncertainty HU is high. It is worth noting that HU values obtained from DL models are generally higher than the counterparts for ML models, reaching a maximum higher than 20% (40% if considering the standard deviation) for Sample_14. In

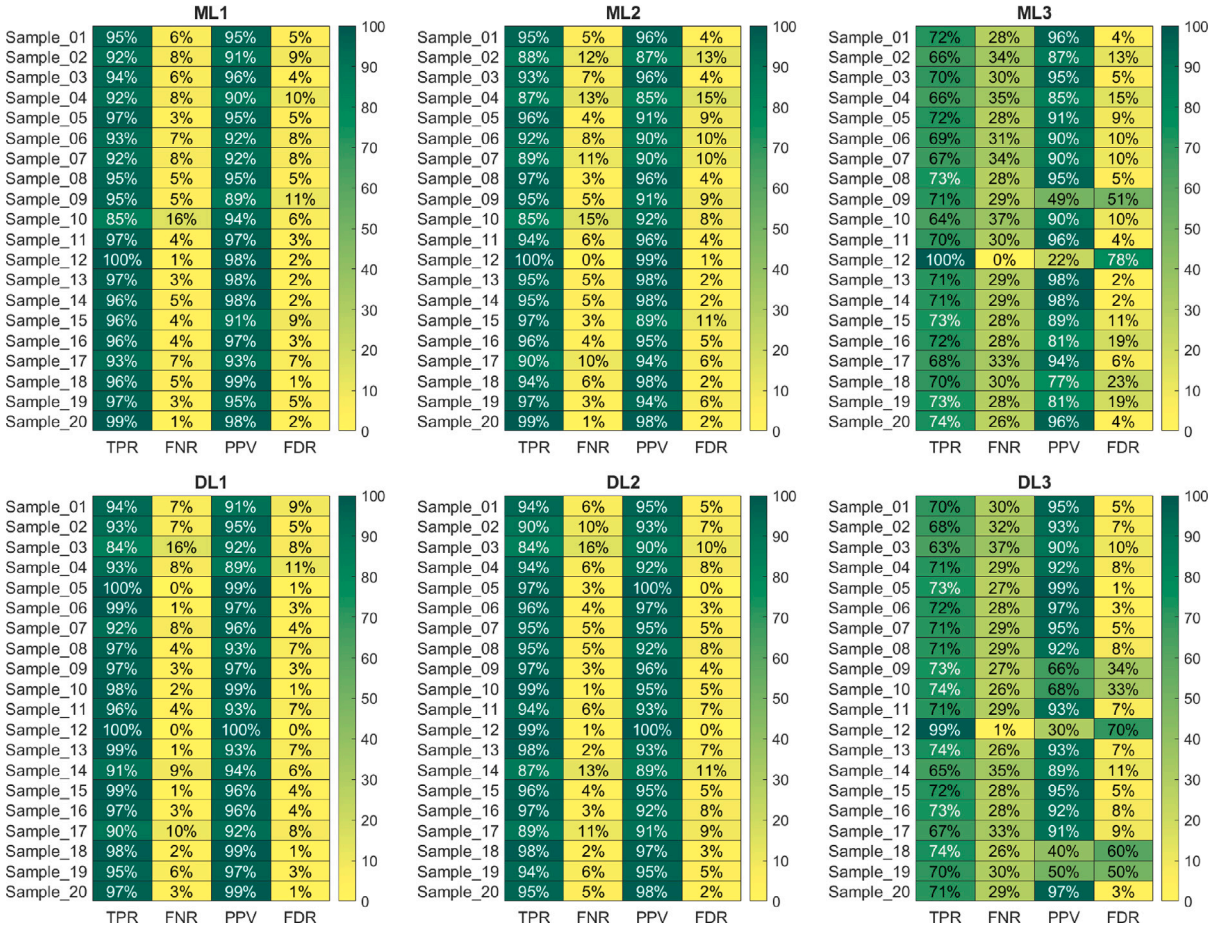

Fig. 6. Summary of performance metrics computed on the test datasets of the corresponding ML and DL models. Results are summarized per sample ID.

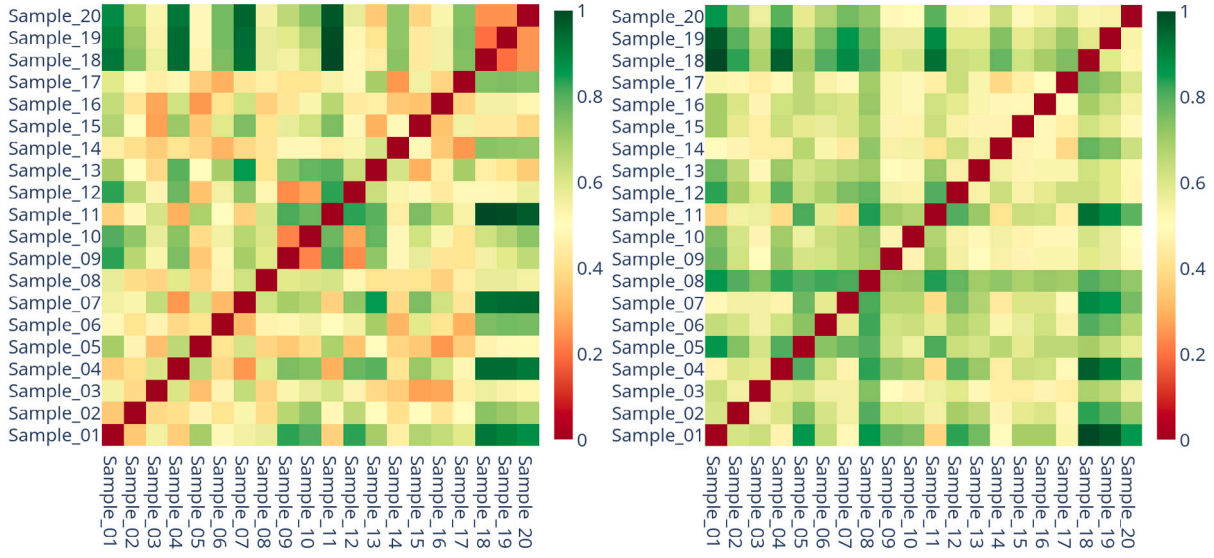

Fig. 7. Energy matrices related to (a) test dataset "α" used for ML1 ($\mathbf{H}_\alpha$), and (b) test dataset "γ" used for DL1 ($\mathbf{H}_\gamma$). Values range from 0 (red) to 1 (green), indicating pairs of classes with very similar or very distant PDFs, respectively.

comparison, for ML models, the average uncertainty is of 10% (20% if considering the standard deviation). This result can be explained by the Energy matrix $\mathbf{H}_\gamma$ in Fig. 7(b) in which most classes have a value of 0.5, thus the formula applies a higher penalty. In fact, the Energy matrix $\mathbf{H}_\gamma$ contains more values closer to 1 (distant case, resulting in heavier penalties in the formulation of HU) compared to $\mathbf{H}_\alpha$, partially explaining why for DL models HU values are higher and with higher variability than the ones obtained for ML models. This behavior reflects the tendency of DL architectures to confuse classes regardless of their similarity (e.g., Energy) more often than ML models.

Finally, for ML3 and DL3, the confusion matrices highlight how the models typically predict almost all classes as Sample_12. However, the corresponding uncertainty metrics are low, suggesting that the models are always confident of their predictions despite being wrong 25% of the time (100 out of 400 occurrences per class) for almost all the classes. This behavior is extremely dangerous and highlights that ML3 and DL3 are not robust toward unseen data. This reflects the difficulty of the problem of raw milk sample classification, already demonstrated by the clustering results in Fig. 4.

To conclude, ML and DL models demonstrated a very similar behavior, reaching the same prediction accuracy overall. The only difference between the two approaches is in the uncertainty results, which may help users choose the right approach for their specific application. By coupling geometry-based and homophily-based uncertainties, it is possible to deeply understand the models' behavior. On one hand, DL models confuse distant classes more often and are sure of the predicted class, making it harder for the final user to understand why misclassification happens. On the other hand, since ML models are feature-based, it is easier to understand why certain classes get confused with each other, given the more intuitive explainability of the features and their relationship with the SP formation process. Higher geometry-based uncertainty values coupled with low homophily-based uncertainty values suggest that the models correctly distinguish among classes given their features, but give room to prediction uncertainty for all the cases in which misclassification happen. This behavior can be a plus if more refined ML models were to be developed that incorporate the "uncertain case".

5. Conclusions

In this work, we have presented the preliminary results proving the applicability of an innovative measurement method based on the integration of SP imaging and artificial intelligence models (both ML and DL) to carry out analyses on raw cow milk samples, for the first time to the best of our knowledge. Using a low-cost, easy-to-use instrumental configuration featuring a red-emitting semiconductor laser and a CMOS camera, we have carried out 4 experimental Campaigns and collected a total of 24,000 SP images generated by illuminating 20 milk samples with similar nutritional composition. To understand the applicability of the method, we tested one off-the-shelf ML model and one DL model developed for milk SP analysis. The goal of our study was to investigate which strategy is more promising and why when applied to milk samples with similar nutritional composition. In addition, the impact of the samples' temperature was investigated as well. To this aim, an important part of our work has been devoted to the metrology aspect of the presented methodology, that is, the estimation of the prediction uncertainty. In particular, uncertainty analysis tries to explain how the similarity of the samples is reflected by the models' performance. By combining all the available metrics (confusion matrices, performance metrics, uncertainty metrics), the performance of the tested models can be comprehended in full.

In conclusion, SP combined with intelligent methods to analyze the results demonstrated to be a powerful tool capable of successfully distinguishing between milk samples of raw cow milk with similar nutritional content, especially if the analysis is coupled with all the relevant metrics to understand the models' behavior. Both feature-based and image-based approaches were successful, achieving an overall accuracy higher than 90%. Compared with ML models, uncertainty metrics highlighted that DL models tend to be more certain of their predictions, even if they confuse both close and distant classes equally, not giving room for doubt or explainability of the results. This is a characteristic reflecting the models' architecture and serves as a guide for the final users in choosing the right model for their application. Moreover, as highlighted by the cluster-based results of the data, temperature has a great impact on SP images formation, thus producing higher uncertainty in those models that were not trained to see the data acquired at different temperatures. In fact, measurement

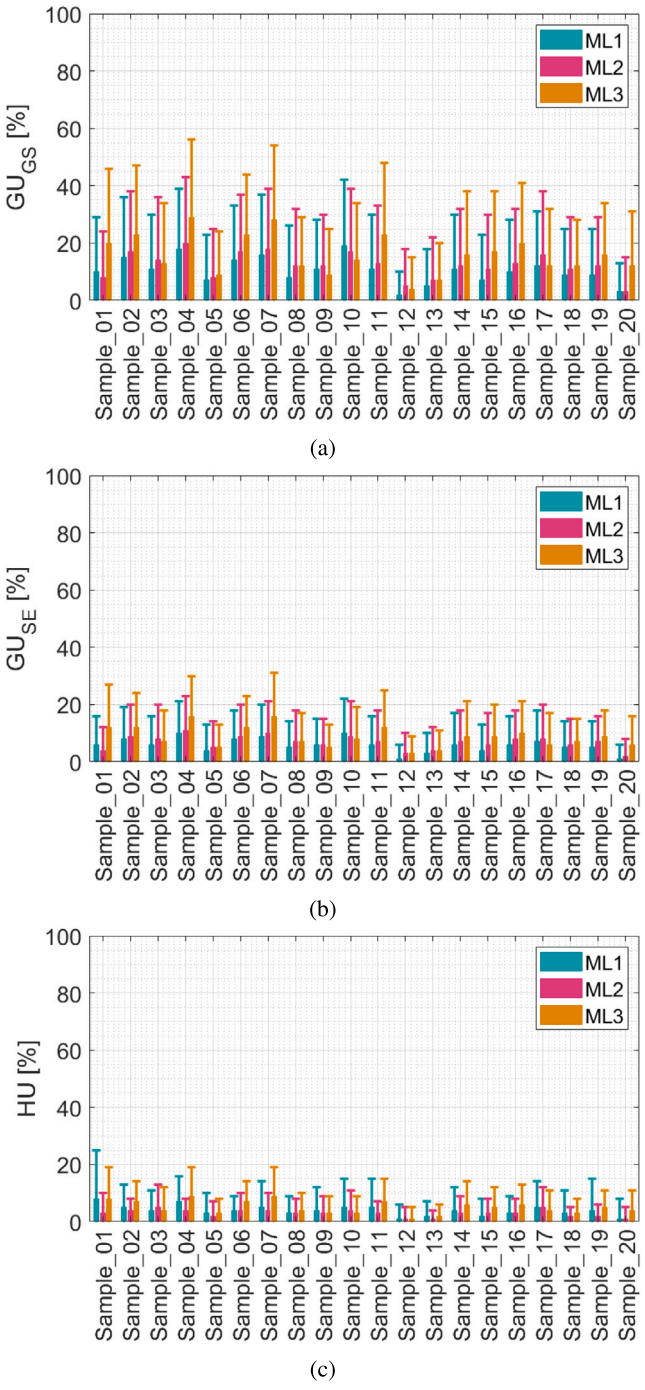

Fig. 8. Uncertainty values for machine learning models ML1–ML3 computed for the three metrics, (a) GU_{GS}, (b) GU_{SE}, and (c) HU.

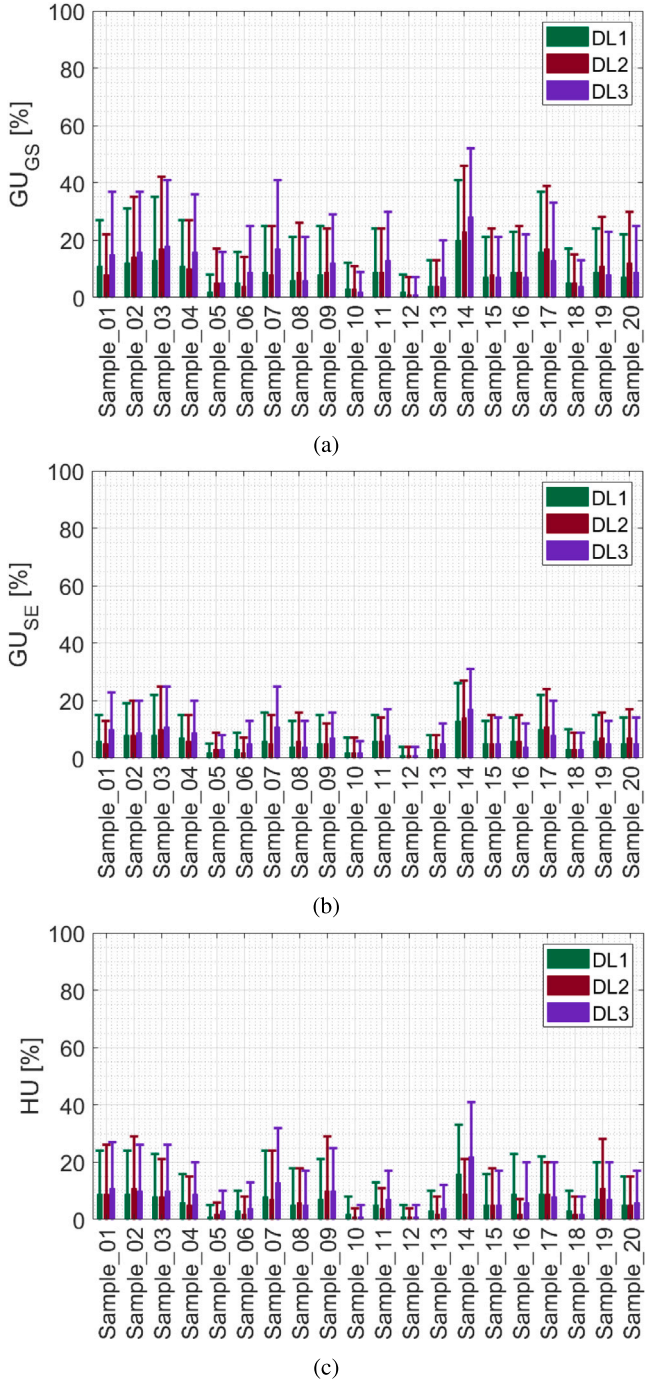

Fig. 9. Uncertainty values for deep learning models DL1–DL3 computed for the three metrics, (a) GU_{GS}, (b) GU_{SE}, and (c) HU.

conducted at low temperatures resulted in better separated clusters. However, conducting measurements on refrigerated raw cow milk is not often possible (e.g., after milking in stables), so we demonstrated the applicability of the method also on samples at ambient temperature.

Given the success of the conducted analysis, future works will be devoted to improving the instrumental configuration, focusing on commercial milk samples that are properly pasteurized. New ML and DL models will also be considered for the analysis, thus advancing the state-of-the-art of measurements for milk quality. As a future development, we will also expand our dataset by including more cow breeds, farming conditions, and seasonal variations, and we will focus on grouping different cows based on specific parameters (e.g., nutritional content, lactation stage, dietary regime). In addition, the reflectance properties of milk captured by hyperspectral imaging techniques can be used as well as input features for a classifier, as demonstrated by [72]; hence, this approach will be investigated in the future in comparison with the SP-based technique presented in this work. Finally, despite being the stepping stone toward a full comprehension of the measurement pipeline proposed, we only scratched the surface of the problem. A deeper study on how lipids, proteins, and lactose affect the generation of SP images is still lacking. Leveraging the findings described in this article, our research will focus on this topic in the near future.

CRediT authorship contribution statement

Cristina Nuzzi: Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation. **Simone Pasinetti:** Writing – review & editing, Validation, Software, Project administration, Methodology, Funding acquisition, Data curation. **Irene Bassi:** Writing – review & editing, Resources, Investigation, Formal analysis. **Valentina Bello:** Writing – original draft, Resources, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This project was supported by the European Union program “Next Generation EU, M4C2 component, investment 1.1”, issued as an open call for the national project “PRIN PNRR 2022”, grant agreement N°P2022FX228 - MUSKETEER: Milk adUlteration detection using SpecKlE paTtern and machinE lEaRning.

The authors thank Matteo Fiocchi (University of Pavia), Francesca Cattivelli (University of Brescia), Paolo Bellassi, and Luca Cattaneo (Universitá Cattolica del Sacro Cuore) for their help in conducting this research. The authors also thank Prof. Sabina Merlo (University of Pavia) for her fruitful suggestions.

Data availability

Data will be made available on request.

References

[1] CLAL, Milk - balance sheet in mio tons | year 2024, 2025, https://www.clal.it/en/?section=home. (Accessed 20 January 2025).
[2] CLAL, EU-27: milk production and population, 2025, https://www.clal.it/en/?section=produzioni_popolazione. (Accessed 20 January 2025).
[3] E. Vanbergue, L. Delaby, J.-L. Peyraud, S. Colette, Y. Gallard, C. Hurtaud, Effects of breed, feeding system, and lactation stage on milk fat characteristics and spontaneous lipolysis in dairy cows, J. Dairy Sci. 100 (6) (2017) 4623–4636.
[4] C. Becker, R. Collier, A. Stone, Invited review: Physiological and behavioral effects of heat stress in dairy cows, J. Dairy Sci. 103 (8) (2020) 6751–6770.
[5] M. Brandt, A. Haeussermann, E. Hartung, Invited review: Technical solutions for analysis of milk constituents and abnormal milk, J. Dairy Sci. 93 (2) (2010) 427–436.
[6] U. Das, R. Biswas, Developing and implementing a facile colorimetric method for detecting salicylic acid in raw milk via biogenic plasmonic nanostructures, Measurement 242 (2025) 115818.
[7] A.R. Sadrolhosseini, S.M. Hamidi, Y. Mazhdi, Detection of gentamicin in water and milk using chitosan-ZnS-Au nanocomposite based on surface plasmon resonance imaging sensor, Measurement 239 (2025) 115412.
[8] M. Werteker, S. Huber, S. Kuchling, B. Rossmann, M. Schreiner, Differentiation of milk by fatty acid spectra and principal component analysis, Measurement 98 (2017) 311–320.

[9] International Organization for Standardization, ISO 19662:2018 | IDF 238:2018, 2018, https://www.iso.org/standard/65935.html. (Accessed 17 January 2025).
[10] International Organization for Standardization, ISO 23318:2022 | IDF 249:2022, 2022, https://www.iso.org/standard/75226.html. (Accessed 17 January 2025).
[11] International Organization for Standardization, ISO 23318:2022 | IDF 249:2022, 2009, https://www.iso.org/standard/41320.html. (Accessed 17 January 2025).
[12] International Organization for Standardization, ISO 22662:2024 | IDF 198:2024, 2024, https://www.iso.org/standard/84827.html. (Accessed 20 January 2025).
[13] H. Soyeurt, P. Dardenne, F. Dehareng, G. Lognay, D. Veselko, M. Marlier, C. Bertozzi, P. Mayeres, N. Gengler, Estimating fatty acid content in cow milk using mid-infrared spectrometry, J. Dairy Sci. 89 (9) (2006) 3690–3695.
[14] Y. Etzion, R. Linker, U. Cogan, I. Shmulevich, Determination of protein concentration in raw milk by mid-infrared Fourier transform infrared/attenuated total reflectance spectroscopy, J. Dairy Sci. 87 (9) (2004) 2779–2788.
[15] R. Linker, Y. Etzion, Potential and limitation of mid-infrared attenuated total reflectance spectroscopy for real time analysis of raw milk in milking lines, J. Dairy Res. 76 (1) (2009) 42–48.
[16] M. De Marchi, V. Toffanin, M. Cassandro, M. Penasa, Invited review: Mid-infrared spectroscopy as phenotyping tool for milk traits, J. Dairy Sci. 97 (3) (2014) 1171–1186.
[17] Z. Schmilovitch, I. Shmulevich, A. Notea, E. Maltz, Near infrared spectrometry of milk in its heterogeneous state, Comput. Electron. Agric. 29 (3) (2000) 195–207.
[18] T. Woodcock, G. Downey, C.P. O'Donnell, Better quality food and beverages: the role of near infrared spectroscopy, J. Near Infrared Spectros. 16 (1) (2008) 1–29.
[19] H.G. Yakubu, Z. Kovacs, T. Toth, G. Bazar, The recent advances of near-infrared spectroscopy in dairy production—A review, Crit. Rev. Food Sci. Nutr. 62 (3) (2022) 810–831.
[20] N. Liu, H.A. Parra, A. Pustjens, K. Hettinga, P. Mongondry, S.M. Van Ruth, Evaluation of portable near-infrared spectroscopy for organic milk authentication, Talanta 184 (2018) 128–135.
[21] M. Kawasaki, S. Kawamura, M. Tsukahara, S. Morita, M. Komiya, M. Natsuga, Near-infrared spectroscopic sensing system for on-line milk quality assessment in a milking robot, Comput. Electron. Agric. 63 (1) (2008) 22–27.
[22] R. Tsenkova, S. Atanassova, K. Itoh, Y. Ozaki, K. Toyoda, Near infrared spectroscopy for biomonitoring: cow milk composition measurement in a spectral region from 1,100 to 2,400 nanometers, J. Anim. Sci. 78 (3) (2000) 515–522.
[23] A. Trani, G. Gambacorta, P. Loizzo, A. Cassone, C. Fasciano, A. Zambrini, M. Faccia, Comparison of HPLC-RI, LC/MS-MS and enzymatic assays for the analysis of residual lactose in lactose-free milk, Food Chem. 233 (2017) 385–390.
[24] D. Gille, B. Walther, R. Badertscher, A. Bosshart, C. Brügger, M. Brühlhart, R. Gauch, P. Noth, G. Vergères, L. Egger, Detection of lactose in products with low lactose content, Int. Dairy J. 83 (2018) 17–19.
[25] A.B. Nongonierma, R.J. FitzGerald, Enhancing bioactive peptide release and identification using targeted enzymatic hydrolysis of milk proteins, Anal. Bioanal. Chem. 410 (2018) 3407–3423.
[26] H.E. Indyk, S. Hart, T. Meerkerk, B.D. Gill, D.C. Woollard, The β-lactoglobulin content of bovine milk: Development and application of a biosensor immunoassay, Int. Dairy J. 73 (2017) 68–73.
[27] A. Poghossian, H. Geissler, M.J. Schöning, Rapid methods and sensors for milk quality monitoring and spoilage detection, Biosens. Bioelectron. 140 (2019) 111272.
[28] T.-F. Cho, A. Yassoralipour, Y.-Y. Lee, T.-K. Tang, O.-M. Lai, L.-C. Chong, C.-H. Kuan, E.-T. Phuah, Evaluation of milk deterioration using simple biosensor, J. Food Meas. Charact. (2021) 1–11.
[29] R.S. Lima, G.C. Danielski, A.C.S. Pires, Mastitis detection and prediction of milk composition using gas sensor and electrical conductivity, Food Bioprocess Technol. 11 (2018) 551–560.
[30] B. Yuan, H. Nørstebø, A.C. Whist, N. Belbachir, Detection of lameness and mastitis pathogens in milk using visual and olfactory sensing, 2020.
[31] R.-S. Lu, G.-Y. Tian, D. Gledhill, S. Ward, Grinding surface roughness measurement based on the co-occurrence matrix of speckle pattern texture, Appl. Opt. 45 (35) (2006) 8839–8847.
[32] S.C. Schneider, S.J. Rupitsch, B.G. Zagar, Signal processing for laser-speckle strain-measurement techniques, IEEE Trans. Instrum. Meas. 56 (6) (2007) 2681–2687.
[33] Y. Zhou, Q. Zuo, L. Zhou, B. Yang, Z. Liu, Y. Liu, L. Tang, S. Dong, Z. Jiang, Image feature based quality assessment of speckle patterns for digital image correlation measurement, Measurement 222 (2023) 113590.
[34] R. Paris, M. Melik-Merkumians, G. Schitter, Probabilistic absolute position sensor based on objective laser speckles, IEEE Trans. Instrum. Meas. 65 (5) (2016) 1188–1196.
[35] D.A. Boas, A.K. Dunn, Laser speckle contrast imaging in biomedical optics, J. Biomed. Opt. 15 (1) (2010) 011109–011109.
[36] Q. Zhang, J.C. Gamekkanda, A. Pandit, W. Tang, C. Papageorgiou, C. Mitchell, Y. Yang, M. Schwaerzler, T. Oyetunde, R.D. Braatz, et al., Extracting particle size distribution from laser speckle with a physics-enhanced autocorrelation-based estimator (PEACE), Nat. Commun. 14 (1) (2023) 1159.
[37] D. Héran, M. Ryckewaert, Y. Abautret, M. Zerrad, C. Amra, R. Bendoula, Combining light polarization and speckle measurements with multivariate analysis to predict bulk optical properties of turbid media, Appl. Opt. 58 (30) (2019) 8247–8256.
[38] H. Loutfi, F. Pellen, B. Le Jeune, G. Le Brun, M. Abboud, Polarized laser speckle images produced by calibrated polystyrene microspheres suspensions: comparison between backscattering and transmission experimental configurations, Laser Phys. 33 (8) (2023) 086001.
[39] V. Bello, L. Coghe, A. Gerbasi, E. Figus, A. Dagliati, S. Merlo, Machine learning-based approach towards identification of pharmaceutical suspensions exploiting speckle pattern images, Sensors 24 (20) (2024) 6635.
[40] F. Vernuccio, A. Bresci, V. Cimini, A. Giuseppi, G. Cerullo, D. Polli, C.M. Valensise, Artificial intelligence in classical and quantum photonics, Laser Photonics Rev. 16 (5) (2022) 2100399.
[41] T. Jakubczyk, D. Jakubczyk, A. Stachurski, Assessing the properties of a colloidal suspension with the aid of deep learning, J. Quant. Spectrosc. Radiat. Transfer 261 (2021) 107496.
[42] J. Yan, M. Jin, Z. Xu, L. Chen, Z. Zhu, H. Zhang, Recognition of suspension liquid based on speckle patterns using deep learning, IEEE Photonics J. 13 (1) (2020) 1–7.
[43] D. Endo, T. Kono, Y. Koike, H. Kadono, J. Yamada, U.M. Rajagopalan, Application of laser speckles and deep learning in discriminating between the size and concentrations of supermicroplastics, Opt. Contin. 1 (11) (2022) 2259–2273.
[44] V. Bello, M. Fiocchi, I. Bassi, E. Figus, S. Merlo, Speckle pattern imaging for recognition of milk dilutions, in: 2024 IEEE International Instrumentation and Measurement Technology Conference, I2MTC, IEEE, 2024, pp. 1–5.
[45] K. Nyandey, D. Jakubczyk, Combining transmission speckle photography and convolutional neural network for determination of fat content in cow milk: an exercise in classification of parameters of a complex suspension, Eur. Phys. J. Plus 139 (2) (2024) 123.
[46] M.-C. Michalski, V. Briard, F. Michel, Optical parameters of milk fat globules for laser light scattering measurements, Le Lait 81 (6) (2001) 787–796.
[47] J.W. Goodman, Speckle Phenomena in Optics: Theory and Applications, second ed., SPIE, 2020, http://dx.doi.org/10.1117/3.2548484.
[48] P. Zajác, J. Čapla, V. Vietoris, S. Zubrická, J. Čurlej, Effects of storage on the major constituents of raw milk, Potravina. Slovak J. Food Sci. 9 (1) (2015) 375–381.
[49] D.M. Amaral-Phillips, Dairy feeding and management considerations during heat stress, 2024, Online. https://afs.ca.uky.edu/content/dairy-feeding-and-management-considerations-during-heat-stress. (Accessed 3 April 2025).
[50] S. Bianchi, R. Pruner, G. Vizsnyiczai, C. Maggi, R. Di Leonardo, Active dynamics of colloidal particles in time-varying laser speckle patterns, Sci. Rep. 6 (1) (2016) 27681.
[51] A. Gastélum-Barrios, G.M. Soto-Zarazúa, A. Escamilla-García, M. Toledano-Ayala, G. Macías-Bobadilla, D. Jauregui-Vazquez, Optical methods based on ultraviolet, visible, and near-infrared spectra to estimate fat and protein in raw milk: A review, Sensors 20 (12) (2020) 3356.
[52] Y. Piederrière, J. Cariou, Y. Guern, B. Le Jeune, G. Le Brun, J. Lotrian, Scattering through fluids: speckle size measurement and Monte Carlo simulations close to and into the multiple scattering, Opt. Express 12 (1) (2004) 176–188.
[53] Y. Piederrière, F. Boulvert, J. Cariou, B.L. Jeune, Y. Guern, G.L. Brun, Backscattered speckle size as a function of polarization: influence of particle-size and-concentration, Opt. Express 13 (13) (2005) 5030–5039.
[54] V. Bello, E. Bodo, S. Merlo, Speckle pattern acquisition and statistical processing for analysis of turbid liquids, IEEE Trans. Instrum. Meas. 72 (2023) 1–4.
[55] V. Bello, I. Bassi, M. Fiocchi, C. Nuzzi, P. Bellassi, L. Cattaneo, S. Pasinetti, Milky WAI: Unlocking the secrets of raw cow milk through speckle pattern and AI, in: 2024 IEEE International Conference on Metrology for EXtended Reality, Artificial Intelligence and Neural Engineering, MetroXRAINE, 2024.
[56] J.C. Dainty, J.W. Goodman, G. Parry, T.S. McKechnie, M. Françon, A.E. Ennos, Laser Speckle and Related Phenomena, Springer Berlin, Heidelberg, 1984, http://dx.doi.org/10.1007/978-3-662-43205-1.
[57] R. Pandiselvam, V. Mayookha, A. Kothakota, S. Ramesh, R. Thirumdas, P. Juvvi, Biospeckle laser technique – A novel non-destructive approach for food quality and safety detection, Trends Food Sci. Technol. 97 (2020) 1–13.
[58] R.M. Haralick, K. Shanmugam, I. Dinstein, Textural features for image classification, IEEE Trans. Syst. Man Cybern. SMC-3 (6) (1973) 610–621.
[59] M.S. Nixon, A.S. Aguado, Feature Extraction & Image Processing for Computer Vision, third ed., Academic Press, 2013.
[60] N. Otsu, A threshold selection method from gray-level histograms, IEEE Trans. Syst. Man Cybern. 9 (1) (1979) 62–66.
[61] E. Kreyszig, Advanced Engineering Mathematics, fourth ed., Wiley, 1979.
[62] N. Altman, M. Krzywinski, Ensemble methods: bagging and random forests, Nature Methods 14 (2017) 933–934.
[63] K. Goyal, P. Kumar, K. Verma, XAI-empowered IoT multi-sensor system for real-time milk adulteration detection, Food Control 164 (2024) 110495.

[64] H.A. Neto, W.L. Tavares, D.C. Ribeiro, R.C. Alves, L.M. Fonseca, S.V. Campos, On the utilization of deep and ensemble learning to detect milk adulteration, BioData Min. 12 (13) (2019).
[65] A. Fornaser, M. De Cecco, P. Bosetti, T. Mizumoto, K. Yasumoto, Sigma-z random forest, classification and confidence, Meas. Sci. Technol. 30 (2) (2018) 025002.
[66] S. Chlaily, D. Ratha, P. Lozou, A. Marinoni, On measures of uncertainty in classification, IEEE Trans. Signal Process. 71 (2023) 3710–3725.
[67] C.M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics), Springer New York, NY, 2006.
[68] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (3) (1948) 379–423.
[69] M.L. Rizzo, G.J. Székely, Energy distance, WIREs Comput. Stat. 8 (1) (2016) 27–38.
[70] P. Virtanen, R. Gommers, T. Oliphant, et al., SciPy 1.0: fundamental algorithms for scientific computing in Python, Nature Methods 17 (2020) 261–272.
[71] I.T. Jolliffe, J. Cadima, Principal component analysis: a review and recent developments, Philos. Trans. Ser. A Math. Phys. Eng. Sci. 374 (2065) (2016) 20150202.
[72] M. Aqeel, A. Sohaib, M. Iqbal, S.S. Ullah, Milk adulteration identification using hyperspectral imaging and machine learning, J. Dairy Sci. 108 (2) (2025) 1301–1314.