Computers in Industry 164 (2025) 104173

Contents lists available at ScienceDirect

Computers in Industry

journal homepage: www.sciencedirect.com/journal/computers-in-industry

Detecting coagulation time in cheese making by means of computer vision and machine learning techniques

Andrea Loddo [a,*], Cecilia Di Ruberto [a], Giuliano Armano [a], Andrea Manconi [b]

[a] Department of Mathematics and Computer Science, University of Cagliari, 09124 Cagliari, Italy
[b] Institute of Biomedical Technologies, National Research Council, CNR, via Fratelli Cervi 93, Segrate, 20054, Milan, Italy

ARTICLE INFO

Keywords:
Image processing
Computer vision
Machine Learning
Food industry
Curd-firming time detection

ABSTRACT

Cheese production, a globally cherished culinary tradition, faces challenges in ensuring consistent product quality and production efficiency. The critical phase of determining cutting time during curd formation significantly influences cheese quality and yield. Traditional methods often struggle to address variability in coagulation conditions, particularly in small-scale factories. In this paper, we present several key practical contributions to the field, including the introduction of CM-IDB, the first publicly available image dataset related to the cheese-making process. Also, we propose an innovative artificial intelligence-based approach to automate the detection of curd-firming time during cheese production using a combination of computer vision and machine learning techniques. The proposed method offers real-time insights into curd firmness, aiding in predicting optimal cutting times. Experimental results show the effectiveness of integrating sequence information with single image features, leading to improved classification performance. In particular, deep learning-based features demonstrate excellent classification capability when integrated with sequence information. The study suggests the suitability of the proposed approach for integration into real-time systems, especially within dairy production, to enhance product quality and production efficiency.

1. Introduction

Cheese dairy products have a high commercial value in the food industry, also considering that they are a source of proteins, calcium, and micro-nutrients with beneficial effects on bone and muscle health. In addition, thanks to their probiotic content, they increase the health of the digestive tract and positively influence the microbiome.

As the demand for cheese continues to rise, fueled by increasing competition in the food industry, there is an escalating need to enhance production efficiency and ensure consistent product quality. Integrating advanced process control methods becomes crucial to achieving these objectives, leading to improved product quality, reduced waste, optimized material, reduced energy costs, shorter processing times, and greater process flexibility.

Cheese is the fresh or matured product obtained through milk coagulation, followed by the separation of the liquid and solid components (known as whey and curd, respectively). The curds, essential building blocks in cheese production, undergo further processing to yield the diverse varieties enjoyed worldwide. The cheese-making process involves two primary operations: the formation of a milk gel and the subsequent cutting of the gel into curd grains, facilitating the separation of whey. In the initial phase of gel formation, casein micelles undergo colloidal destabilization due to the chemical modification of the protective k-casein coat by coagulating agents. The subsequent phase involves the aggregation of destabilized casein micelles, forming a gel network that evolves into a solid gel (Arango and Castillo, 2018).

The critical phase of milk coagulation is central to cheese production, representing a standardized yet variable process across manufacturers, ranging from large commercial dairies to small artisanal settings. Precisely determining the cutting time during curd formation emerges as a key challenge, which directly affects the quality and quantity of the final cheese product. In fact, in the process of cheese-making, the curd must be precisely cut at a point when it has achieved adequate firmness to form distinct particles, facilitating the expulsion of whey without causing fragmentation. Consequently, the timing of curd cutting is intentionally delayed beyond the initial gelation stage, and its meticulous selection significantly influences the moisture content, yield, and overall quality of the resulting cheese, along with minimizing whey fat losses. The coagulation process and its associated cutting step represent arguably the least controlled stage in the cheese-making procedure, thereby exerting a substantial impact on the overall cheese yield (Arango and Castillo, 2018). The duration of coagulation time serves as an indicative measure of enzyme activity and effectiveness,

* Corresponding author.
E-mail address: andrea.loddo@unica.it (A. Loddo).

https://doi.org/10.1016/j.compind.2024.104173
Received 19 April 2024; Received in revised form 20 August 2024; Accepted 4 September 2024
0166-3615/© 2024 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

particularly in diverse operational conditions. Moreover, it serves as a crucial reference point for determining the optimal time to cut the curd for whey drainage. Hence, the accurate determination of coagulation time holds paramount importance, as it plays a pivotal role in the precise evaluation of cutting time, a critical factor for maximizing both cheese yield and quality.

This challenge is particularly pronounced in large-scale automated production facilities, where variability in coagulation conditions, processing changes, and the potential for human errors introduce complexities in maintaining accurate control over cutting times (Arango and Castillo, 2018; Gao et al., 2022; Guinee, 2021).

The integration of advanced methodologies, such as computer vision techniques, is essential to address the challenges faced in the cheese-making process and optimize it for enhanced production efficiency and product quality. This aligns with the principles of Industry 4.0, which emphasizes the integration of digital technologies, automation, and data-driven decision-making to improve industrial processes (Peres et al., 2018; Bellavista et al., 2023).

The adoption of computer vision techniques (Alarcon and Shene, 2021; Sabzi and Arribas, 2018), as demonstrated in other key industrial scenarios (Kamm et al., 2023; Tchuente et al., 2024), can play a crucial role in enabling smart manufacturing (Xu et al., 2024) and quality control (Haleem et al., 2021) in the cheese industry (Tufano et al., 2018). For example, computer vision can be employed for real-time monitoring and inline measurement of parameters like milk gel firmness, cutting time of cheese curd, and other critical attributes throughout the production process

Current state-of-the-art approaches in cheese production focus extensively on real-time estimates of curd-firming and the prediction of cutting times (Gao et al., 2022; Guinee, 2021; Feng et al., 2021; Vacca et al., 2020). Attention has been directed toward understanding the milk coagulation process, emphasizing the importance of monitoring changes in milk composition and coagulation conditions (Gao et al., 2022; Guinee, 2021; Feng et al., 2021; Lazouskaya et al., 2021). However, existing methodologies often fall short of addressing the nuanced variability introduced by small-scale factories, where flexible production schedules contribute to increased uncertainty in coagulation processes. Large-scale facilities, while highly automated, still face the persistent risk of unforeseen changes in processing conditions, potentially resulting in substantial economic losses.

In response to the outlined challenges, this paper presents a novel, non-invasive, automated approach aimed at detecting curd-firming time from images. It also aims to open the field to improved production efficiency, enhance product quality, and mitigate possible economic losses associated with variations in coagulation conditions. Leveraging machine learning (ML) and computer vision (CV) techniques, the proposed method offers a novel approach to monitoring the curd formation process, providing real-time insights into curd firmness and predicting optimal cutting times. The contributions of this paper can be summarized as follows:

- **Public dataset release.** We hereby provide public access to the first dataset encompassing images of the cheese-making process. It is available via the official FigShare repository.
- **Machine learning-based approach.** We introduce an innovative automated artificial intelligence-based method for detecting curd-firming time from images, employing machine learning and computer vision techniques.
- **Hybrid approach.** The proposed solution leverages both image features and characteristics derived from the temporal relationships among images, enhancing the accuracy of detection.
- **Automated curd-firming time detection.** The proposed solution is particularly significant as precise cutting time influences the quality and quantity of the final cheese product, addressing a critical challenge in the cheese-making process.
- **Consolidated baseline.** By providing a non-invasive and technologically advanced solution, the paper aims to offer a consolidated baseline in the context of curd-firming time identification within the employed dataset.

The structure of this article is organized as follows. Section 2 provides a comprehensive review of existing methodologies employed in the analysis of milk coagulation. This review establishes a foundational context for the proposed approach. Section 3 elucidates details regarding CM-IDB, feature extraction methodologies, evaluation measures, and the ML classifiers adopted, thereby laying the essential groundwork for the study. In Section 4, the paper introduces the underlying framework proposed for this study, delineating its principal focal points. Section 5 delves into the experimental evaluation conducted, offering a presentation of the undertaken experiments along with the corresponding results and subsequent discussions. A more general discussion overview is provided in Section 6. The concluding remarks of this study, along with insightful suggestions for potential enhancements and avenues for future research based on our findings, are given in Section 7.

2. Related work

In recent decades, various methodologies have been employed to monitor milk coagulation for the determination of coagulation and cutting times. These approaches encompass electrical (Hwang et al., 2022), thermal (Feng et al., 2021), optical (Hass et al., 2015; Tabayehnejad et al., 2012), viscometric (Gao et al., 2022; Vacca et al., 2020), and ultrasonic methods (Lazouskaya et al., 2021; Budelli et al., 2017).

CV and digital image analysis serve as effective and non-invasive techniques for probing the optical properties of cheese, providing insights into its composition and structure. These methods have been successfully employed to evaluate various external quality properties, such as color, defects like gas holes or rind defects, and meltability. Image analysis has been applied to measure gas production, identify abnormal shapes or distributions of eyes in cheeses like Emmental and Tilsit (Budžaki et al., 2014), detect rind halo cheese defects, and quantify the area occupied by calcium lactate crystals on naturally smoked Cheddar cheese surfaces (Galli et al., 2023). Moreover, CV and image analysis have enhanced empirical methods (like the Arnott and Schreiber test), offering a new approach for evaluating meltability and oiling-off of Mozzarella cheese (Moghiseh et al., 2021). Digital image analysis has also been instrumental in assessing ingredient additions, their distribution, and for cheese quality evaluation overall (Bosakova-Ardenska, 2024). Scanning Electron Microscopy, combined with image analysis, has been applied to evaluate texture in Sardo cheese ripened at different stages, as well as to assess various freeze-drying cycles during ripening (Pieniazek and Messina, 2020). Several artificial neural network (ANN) models, including single and multilayers, have been developed and compared for predicting the shelf life of processed cheese (Taneja et al., 2023). Near-infrared hyperspectral (NIR-HS) images acquired during ripening have been employed to predict cheese maturity, offering an alternative to subjective evaluation techniques. These images revealed an increasing homogeneity of the cheese over storage and ripening, suggesting maturation initiates at the center and progresses outward (Priyashantha et al., 2020). In a recent study (Loddo et al., 2022), we proposed a non-invasive methodology using CV and ML to monitor the cheese ripening process and automatically detect maturation stages. This approach analyzes images captured with an ordinary photo camera and was specifically applied to a typical soft cheese (*Pecorino*) produced by a Sardinian dairy company. Table 1 summarizes the existing methods with comprehensive details regarding their applications and key features.

In contrast to existing studies that primarily focus on various methods for monitoring milk coagulation and cheese quality, the proposed work introduces a novel and non-invasive approach specifically designed for automating the detection of cheese curd from images. While

Table 1
Comparison of existing methodologies for monitoring cheese production and quality.

Study/Methodology	Focus/Applications	Key features
Traditional Methods	Electrical Methods (Hwang et al., 2022)	Monitoring milk coagulation, non-invasive, real-time data acquisition
	Thermal Methods (Feng et al., 2021)	Coagulation and cutting times, temperature-based analysis
	Optical Methods (Hass et al., 2015; Tabayehnejad et al., 2012)	Evaluating cheese properties, insight into composition and structure
	Viscometric Methods (Gao et al., 2022; Vacca et al., 2020)	Measuring viscosity changes, correlation with cheese quality
	Ultrasonic Methods (Lazouskaya et al., 2021; Budelli et al., 2017)	Monitoring milk coagulation, high-frequency sound waves for analysis
CV and ML Techniques	CV and Digital Image Analysis (Loddo et al., 2022)	Identifying cheese ripeness stages, non-invasive
	Image Analysis for Gas Production (Budžaki et al., 2014)	Identifying abnormal shapes in cheeses, quantitative assessment of gas holes
	Scanning Electron Microscopy (Pieniazek and Messina, 2020)	Evaluating texture in cheese, detailed structural analysis
	ANN Models (Taneja et al., 2023)	Predicting shelf life of processed cheese, ML for predictive analytics
	NIR-HS Imaging (Priyashantha et al., 2020)	Predicting cheese maturity, homogeneity assessment during ripening
Proposed Methodology	**Automating curd detection**	**Non-invasive, simple camera setup, focuses on curd-firming time Integrates CV and ML techniques for cheese production**

Table 2
Summary of extracted handcrafted features. The table specifications encompass the reference paper, the number of features provided, and a short description.

Group	#features	Description
CH_5 (Mukundan et al., 2001)	21	Chebyshev 1st-order moments of order 5
CH2_5 (Mukundan et al., 2001)	15	Chebyshev 2nd-order moments of order 5
LM_5 (Teague, 1980)	21	Legendre moments of order 5
HARri (Haralick et al., 1973)	26	Rotation invariant Haralick texture features
LBP_18 (He and Wang, 1990)	36	Local binary patterns (with a radius of 1 and 8 neighbors)

prior research has explored diverse techniques such as electrical, thermal, optical, and ultrasonic methods, our proposed method leverages CV and image analysis. The innovation lies in using a simple setup involving only a camera connected to a computer, offering a practical and accessible solution for the food industry. Unlike studies that primarily address cheese quality parameters or maturation stages, our work focuses exclusively on the critical phase of cheese production — i.e., curd-firming time determination — by integrating advanced CV and ML techniques. This distinction positions our methodology as a pioneering contribution to the field, addressing a crucial aspect of cheese production that has been less explored in the existing literature.

3. Methods and materials

Section 3.1 gives a detailed explanation of the various features used during our research work, which can be divided into two main groups: handcrafted and deep learning-based features. The former has been extracted from the images by means of well-known algorithms (Section 3.1.1), whereas the latter have been obtained from Convolutional Neural Network CNN architectures (Section 3.1.2). Furthermore, Section 3.2, describes the adopted ML classifiers and Section 3.3 briefly recalls the performance measures used to assess classification results. Section 3.4 illustrates the mathematical tools employed to characterize structural changes manifesting during the milk coagulation phase, whereas Section 3.5 details the dataset, including the sequences of images that faithfully reproduce the milk coagulation process. This dataset serves as the basis for assessing and validating the efficacy of the proposed approach.

3.1. Feature extraction

Within this section, we describe the feature extraction procedures adopted in our analytical framework.

3.1.1. Handcrafted features

Handcrafted features in image analysis encompass various techniques for extracting morphological, pixel-level, and textural information from images (see also Table 2 for a summary). The used handcrafted features will be referenced as HC, hereinafter.

Invariant moments. Image moments, representing weighted averages of pixel intensities, serve to extract specific properties for image analysis and pattern recognition. In our study, two types of moments — Legendre and Chebyshev — were employed, each offering distinctive insights.

- **Chebyshev Moments (CH)**. Introduced by Mukundan et al. (2001), these orthogonal moments, derived from Chebyshev polynomials (Di Ruberto et al., 2018), include both first-order (CH) and second-order (CH_2) moments. In our analysis, we computed CH and CH_2 of order 5.
- **Second-order Legendre Moments (LM)**. Proposed by Teague (1980), these moments, derived from Legendre orthogonal polynomials (Teh and Chin, 1988), capture shape and spatial characteristics. We utilized Legendre moments of order 5.

Texture features. Emphasizing fine textures, the following features were evaluated:

- **Rotation-Invariant Haralick Features (HARri)**. Thirteen Haralick features (Haralick et al., 1973), derived from the Gray Level Co-occurrence Matrix (GLCM), were transformed into rotation-invariant features for enhanced robustness (Putzu and Di Ruberto, 2017).
- **Local Binary Pattern (LBP)**. Described by He and Wang (1990), LBP characterizes texture and patterns. In our study, the histogram of rotation-invariant LBP (LBP_ri) (Ojala et al., 2002) served as the feature vector.

3.1.2. Deep learning-based features

Deep learning feature extraction, coupled with shallow learning classifiers, has proven to be a successful strategy in enhancing the predictive prowess of conventional deep learning models (LeCun et al., 2015). CNNs excel at capturing global features from images by subjecting the input to multiple convolutional filters and progressively reducing dimensionality across various architectural stages. In our experiments, we opted for several pre-trained off-the-shelf architectures based on the Imagenet1k dataset (Deng et al., 2009). A detailed account of the chosen layers for feature extraction, input size, and the number of trainable parameters for each CNN is provided in Table 3. The chosen deep learning-based features will be referenced as DEEP hereinafter.

Table 3
Summary of features extracted from images using CNNs. The table specifications encompass the reference paper, the count of trainable parameters in millions, the input shape, the designated feature extraction layer, and the number of features provided.

Network	Parameters (M)	Input shape	Feature layer	#Features
DarkNet-53 (Redmon et al., 2016)	20.8	224 × 224	Conv53	1000
GoogLeNet (Szegedy et al., 2015)	5	224 × 224	Loss3	1000
Inception-v3 (Szegedy et al., 2016)	21.8	299 × 299	Last FC	1000
ResNet-18 (He et al., 2016)	11.7	224 × 224	Pool5	512
ResNet-101 (He et al., 2016)	44.6	224 × 224	Pool5	1024

3.2. Machine learning techniques

This study employs handcrafted and deep-learning extracted features as inputs for various ML classifiers, including the Random Forest Classifier (RF), k-nearest Neighbor (k-NN), Support Vector Machine (SVM), and Gradient Boosting Classifier (GB). A concise introduction to these classifiers is provided below.

k-Nearest Neighbor (k-NN). The k-NN classifier determines categories based on the classes of the k-training examples nearest in the distance to a given observation. This method adopts a local strategy, considering the proximity of neighboring instances to classify observations effectively.

Support Vector Machine (SVM). This classifier categorizes examples by mapping them to specific sides of a decision boundary. Utilizing a radial basis function kernel to handle non-linear relationships, SVM offers a nuanced representation of complex data patterns. The one-vs-rest approach is employed for multiclass problems, training individual classifiers for each class against the rest.

Random Forest (RF). RF amalgamates predictions from multiple decision trees, each trained on random subsets of features and examples. This ensemble technique enhances model robustness, introducing diversity among trees to improve resilience against data imbalance and mitigate overfitting. The inclusion of 100 trees specifically contributes to strengthening the predictive power of the random forest.

Gradient Boosting Classifier (GB). Operating through a sequential ensemble-building process, the Gradient Boosting Classifier constructs a series of weak learners, typically decision trees. Each successive tree aims to rectify the errors of its predecessor, iteratively refining predictive accuracy. This method excels in capturing intricate relationships within data, making it particularly effective in scenarios with complex and non-linear patterns.

Multi-Layer Perceptron (MLP). It is a type of artificial neural network composed of multiple layers of nodes or neurons. It typically consists of an input layer, one or more hidden layers, and an output layer. Neurons in each layer are interconnected with weighted connections, and each neuron applies an activation function to its weighted inputs. MLPs are trained using supervised learning techniques such as backpropagation (Haykin and Network, 2004).

3.3. Performance evaluation measures

The classification performance has been evaluated in terms of *accuracy*, *specificity*, *sensitivity*, and *F1*. Clear explanations of these measures for binary classification tasks are provided below. To assess a binary classifier's performance on a dataset, each instance within the dataset will be categorized as either negative or positive based on the classifier's predictions. Comparing classification results against true target values determines whether an instance belongs to one of the following elements, which are part of the confusion matrix:

True Negatives (TN). The count of instances from the negative class that have been correctly predicted as negative.

False Positives (FP). The count of instances from the negative class that have been erroneously predicted as positive.

False Negatives (FN). The count of instances from the positive class that have been erroneously predicted as negative.

True Positives (TP). The count of instances from the positive class that have been correctly predicted as positive.

All the aforementioned measures can be derived from the confusion matrix and are defined as follows:

Precision (PRE) is the fraction of positive instances correctly classified among all instances classified as positive.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

Specificity (SPE) is the ratio of correctly predicted negative instances to the total actual negative instances. It focuses on minimizing false positives.

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{2}$$

Sensitivity (SEN) is the ratio of correctly predicted positive instances to the total actual positive instances. It quantifies how well the classifier identifies positive instances and aims to minimize false negatives.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{3}$$

Accuracy measures how often a classifier correctly predicts positive and negative instances. It provides a general overview of the classifier's performance, measuring the overall correctness of predictions. However, due to the high imbalance of CM-IDB, a different definition of accuracy has been adopted in this work, called balanced (or unbiased) accuracy. This measure is obtained by ignoring the ratio between negative and positive instances so that it is, in fact, the mean between specificity and sensitivity.

$$\text{Accuracy} = \frac{SPE + SEN}{2} \tag{4}$$

F1 (also called F-measure or F-score) is the harmonic mean between precision and sensitivity. It provides a balance between the two.

$$\text{F1} = \frac{2 \cdot SPE \cdot SEN}{SPE + SEN} \tag{5}$$

3.4. Measures for image temporal analysis

This section delineates two alternative methodologies, specifically autocorrelation (Section 3.4.1) and structural similarity index (SSIM) (Section 3.4.2), employed for the assessment of structural alterations occurring throughout the coagulation process. These methodologies prove instrumental in identifying the curd-firming time.

3.4.1. Autocorrelation for image sequence analysis

Autocorrelation, a mathematical tool employed for detecting recurrent patterns, finds application in diverse domains, including signal processing, statistics, image processing, and astrophysics. Specifically in signal processing, it is utilized to identify repetitive patterns within a signal over time. Considering images as spatial signals, the assessment of spatial autocorrelation becomes a meaningful endeavor (Ionescu et al., 2018).

The realm of image processing encompasses a variety of techniques grounded in the autocorrelation of images, serving diverse purposes ranging from texture analysis to the estimation of grain density. In the

field of remote sensing, the probability of change, indicative of the likelihood of alterations in a specific area, is often derived through temporal analysis of historical datasets. Autocorrelation emerges as a valuable tool for assessing this change probability (Panuju et al., 2020), as well as for ecosystem evaluation (Kang et al., 2022). Urban areas have also been subject to analysis employing both global and local autocorrelation indices (Yuan et al., 2019). This approach facilitates the depiction of pattern features related to urban expansion, thereby enhancing estimations of land use change.

In a study proposed by Mondini (2017), the authors present a methodology for gauging spatial autocorrelation changes induced by event landslides in a temporal series of synthetic aperture radar intensity *Sentinel-1* images. Autocorrelation, fundamentally designed to gauge the relationship between a variable's present and past values, proves instrumental in change identification.

Spatial autocorrelation, when applied to images, offers insights into their similarity, assessing relationships by evaluating displacements and deformations of image points. Correlation values range from -1 to $+1$, signifying perfect negative and positive correlation, respectively, with 0 indicating no correlation between images. Often, these values are adjusted to the $[0, 1]$ range, where higher values denote greater similarity. Conversely, values near 0 suggest less correlated images.

Examining a series of images depicting the milk coagulation stage in the cheese-making process reveals a visual phenomenon that may aid in identifying or predicting the optimal curd-firming time. Initially, images evolve rapidly, perceptibly changing until reaching the curd-firming time. Subsequently, the images attain homogeneity, exhibiting slow variations. To characterize these temporal changes, the autocorrelation of the images is employed, treating them as a time-ordered data series and analyzing the associated autocorrelation matrix (Ionescu et al., 2018).

3.4.2. Structural similarity index for image comparison

SSIM, serving as a metric, is employed to gauge the similarity between two given images. Introduced in 2004 by Wang et al. (2004), SSIM originally aimed to assess perceptual image quality. Conventional image quality assessment often relies on quantifying errors between a reference and a sample image, typically computed through Mean Squared Error (MSE) by comparing the pixel values of corresponding pixels in the sample and reference images.

However, the human visual perception system excels at identifying structural information in a scene, allowing for the discernment of differences between information extracted from a reference and a sample scene. SSIM endeavors to emulate this behavior and performs well in tasks that involve distinguishing between a sample and a reference image. Unlike other techniques, such as MSE or Peak Signal-to-Noise Ratio, which estimate absolute errors, SSIM considers image degradation as perceived changes in structural information. It incorporates important perceptual phenomena, encompassing luminance and contrast terms. Structural information recognizes that pixels exhibit strong inter-dependencies, especially when spatially close, conveying crucial information about the visual scene's structure. Luminance takes into account that distortions are less visible in bright regions, while contrast acknowledges that distortions become less noticeable in areas with significant activity or texture.

SSIM compares two images based on these extracted features — structure, luminance, and contrast. The resulting SSIM value between two images falls within the range of -1 to $+1$. A value of $+1$ indicates high similarity or identical images, while -1 denotes significant dissimilarity. Often, these values are adjusted to the range $[0, 1]$, where a value of 1 signifies similar images and 0 represents dissimilar images. SSIM finds applications in diverse areas (Bhatt et al., 2021), such as image compression, outperforming MSE (Søgaard et al., 2016); image restoration, where an SSIM variant in Wiener filter implementation yields improved visual results (Wang and Bovik, 2009); and image classification (Moya-Sánchez et al., 2021). In the context of the milk coagulation stage, SSIM serves as a potent tool to characterize structural changes. It can be employed to assess the similarity in the temporal sequence of images depicting the coagulation process.

3.5. Dataset description

In this work, we provide and present the first public dataset encompassing images of the cheese-making process, named the Cheese-Making Image Data Base (CM-IDB).

CM-IDB was built with the support of the Sardinian agency for the implementation of regional agricultural and rural development programs (LAORE[1]) and BiosAbbey S.r.l. In particular, it was curated by gathering images from the Sardinian (Italy) dairy company named "Podda Formaggi".

It is publicly available via a Figshare repository[2] under the CC BY-SA 4.0 license (Loddo, 2024).

Comprising 12 distinct sets of chronologically ordered images, each set captures the coagulation process — depicting the transformation of milk from its initial liquid state to a gelatinous form known as curd. The initiation of this transformation, denoted as the curd-firming time (CF-time), is discerned by an image at varying positions within the series for different sets.

Image acquisition employed a camera equipped with a CMOS 35.9×24.0 mm sensor and a 24 Mpixel resolution. For the sake of completeness the specific camera used was a Nikon D750. All images adhere to the RGB format, possess a resolution of $6,016 \times 4,016$ pixels, and were captured at intervals of approximately 10 s.

Within each set, images are temporally ordered and labeled based on the maturation stage—pre-CF-time (negative class) or CF-time (positive class). CM-IDB exhibits class imbalance, with a majority of images belonging to the negative class. The time interval between subsequent images varies, with negative examples sampled every 10 s and positive examples every 2 s.

Table 4 presents a comprehensive overview of each set, numbered from 1 to 12. The table includes crucial details such as the number of images in each set and the position of the target image within the timing sequence, signifying the CF-time. Notably, Sets 1 and 2 feature images capturing the coagulation process of fresh whole sheep's milk (*Pecorino Romano*), while the subsequent sets involve a mixed composition of cow and sheep milk.

Image acquisition always started one to two minutes after the addition of rennet to the milk, concluding at the coagulation firming stage. The CF-time, manually identified by a dairy expert, is the target image for each set. Fig. 1 illustrates a temporal sequence of images from Set 11, offering visual insight into the coagulation process. To be specific, the sequence begins with 5 images at a temporal interval of 12, followed by the target image, and concludes with 3 images after the target image at a temporal interval of 9 until the sequence's conclusion.

Furthermore, Fig. 2 showcases the target image for each set, emphasizing noticeable visual distinctions among them.

4. The proposed framework

This section provides an overview of the computer vision-based approach employed for estimating the CF-time in the milk coagulation process.

Initially, the images underwent enhancement through adaptive histogram equalization and local filtering, as presented in Section 4.1. Afterward, two distinct approaches have been experimented, called (a) spatial-based and (b) spatiotemporal-based. In the former approach, described in Section 4.2, both HC and DEEP features were extracted from the images and subsequently fed into several ML classifiers. The latter approach described in Section 4.4 integrates spatial features extracted from individual images with temporal ones to achieve the classification objective. Temporal features have been derived from autocorrelation and SSIM measures and used for classification. They

[1] LAORE Sardegna.

[2] CM-IDB official figshare repository.

Table 4
Comprehensive dataset details: the table provides information on the 12 distinct sets of time-ordered images, denoted from 1 to 12. Each set is characterized by its number of images, representing the number of images it contains, and specifies the target image position, crucial for identifying the curd-firming time within the temporal sequence.

Set	1	2	3	4	5	6	7	8	9	10	11	12
Number of images	94	102	128	112	77	70	84	96	105	94	89	111
Target image position	78	91	109	90	55	46	64	73	69	60	61	78

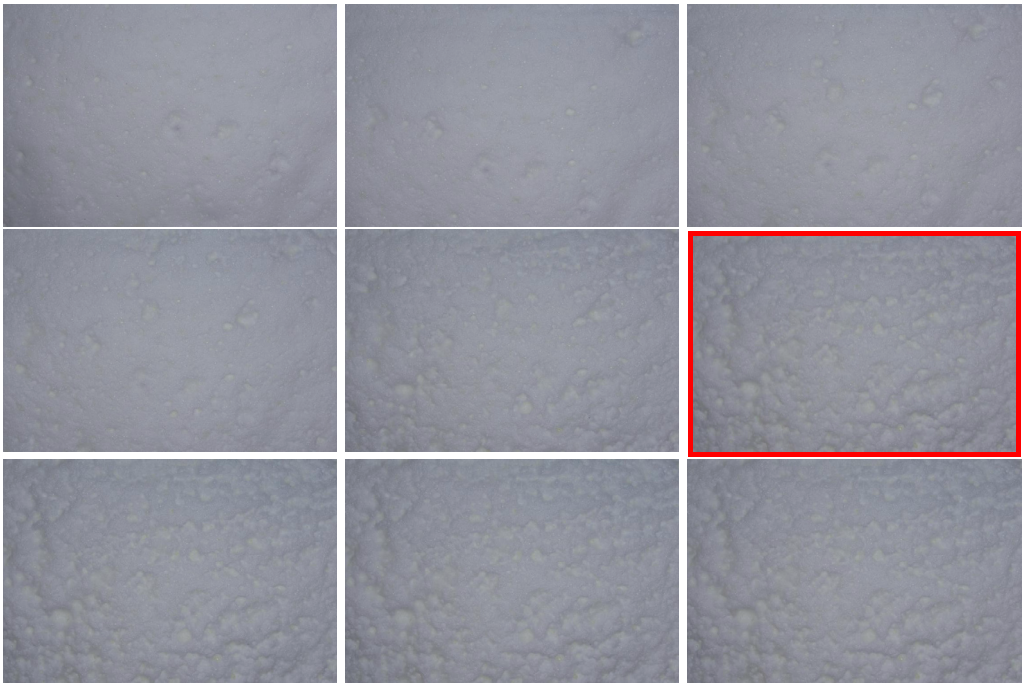


Fig. 1. A sample temporal sequence of 9 images of Set 11: in order, 5 images at a temporal distance equal to 12 from the beginning of the sequence, then the Target image marked by a red box and, finally, 3 images after the Target image at a distance of 9 until the end of the sequence.

were considered with the goal of identifying the CF-time, marking the onset of the milk's transition from a liquid to a gelatinous state, and forming the curd. Fig. 3 provides a schematic representation of the proposed framework.

4.1. Image preprocessing

To enhance local contrast and thereby incorporate additional details conducive to image analysis, a preprocessing technique employing adaptive histogram equalization has been applied to the images. In contrast to regular histogram equalization, this adaptive method computes multiple histograms, each corresponding to a distinct image region, and utilizes them to distribute brightness values. Consequently, this approach effectively reveals details that may not be directly discernible in the original images.

Additionally, to accentuate spatial variations in brightness levels, an enhancement process has been applied to the processed images. Specifically, a filter based on local entropy has been employed. In this context, the brightest pixels in the filtered image correspond to neighborhoods in the original image with higher entropy. Refer to Fig. 4 for a visual representation, including the original image, the enhanced image, and the filtered image.

4.2. Spatial-based approach

The problem at hand is inherently a prediction problem. However, turning prediction problems into classification ones is a common practice, as the latter are typically easier to encode and set up than the former. Hence, the first approach that was experimented with consisted of training models using features extracted from each single image.

The following handcrafted features have been extracted from the images: invariant moments (CH, CH_2, LM) and texture features (HARri and LBP). Some DEEP features have also been considered and extracted from the following CNNs: DarkNet-53, GoogLeNet, Inception-v3, ResNet-18, and ResNet-101. Then, both the HC and DEEP features have been used to train different ML classifiers, including RF, k-NN, SVM, GB, and MLP.

4.3. Structural changes in the timeline of images

Before describing the spatiotemporal-based approach, we believe it is appropriate to delve deeper into the structural changes that occur in the images associated with the coagulation process and how to identify them. Both autocorrelation and SSIM have been evaluated for the temporal series of images. The result is an autocorrelation matrix or an SSIM matrix with normalized correlation coefficients or similarity values between any two-time points of the data. Values close to 1 indicate highly correlated or very similar images, whereas values close to 0 indicate less correlated o very dissimilar images. Fig. 5 shows the autocorrelation matrix and the SSIM matrix for a sample set using a graphical representation. Both cases highlight a significant increase in autocorrelation and similarity over time. A color scheme has been used to emphasize these differences. In particular, negligible changes are marked in yellow, whereas significant ones are marked in blue. In this sample set, the images in the time series are 89, and the image relative to the CF-time is at position 61. The values in the matrices start to visibly increase at that position and remain high till the end of the temporal series. This confirms that both autocorrelation and SSIM measures are useful to characterize the changes happening during the coagulation phase of a cheese-making process. To make

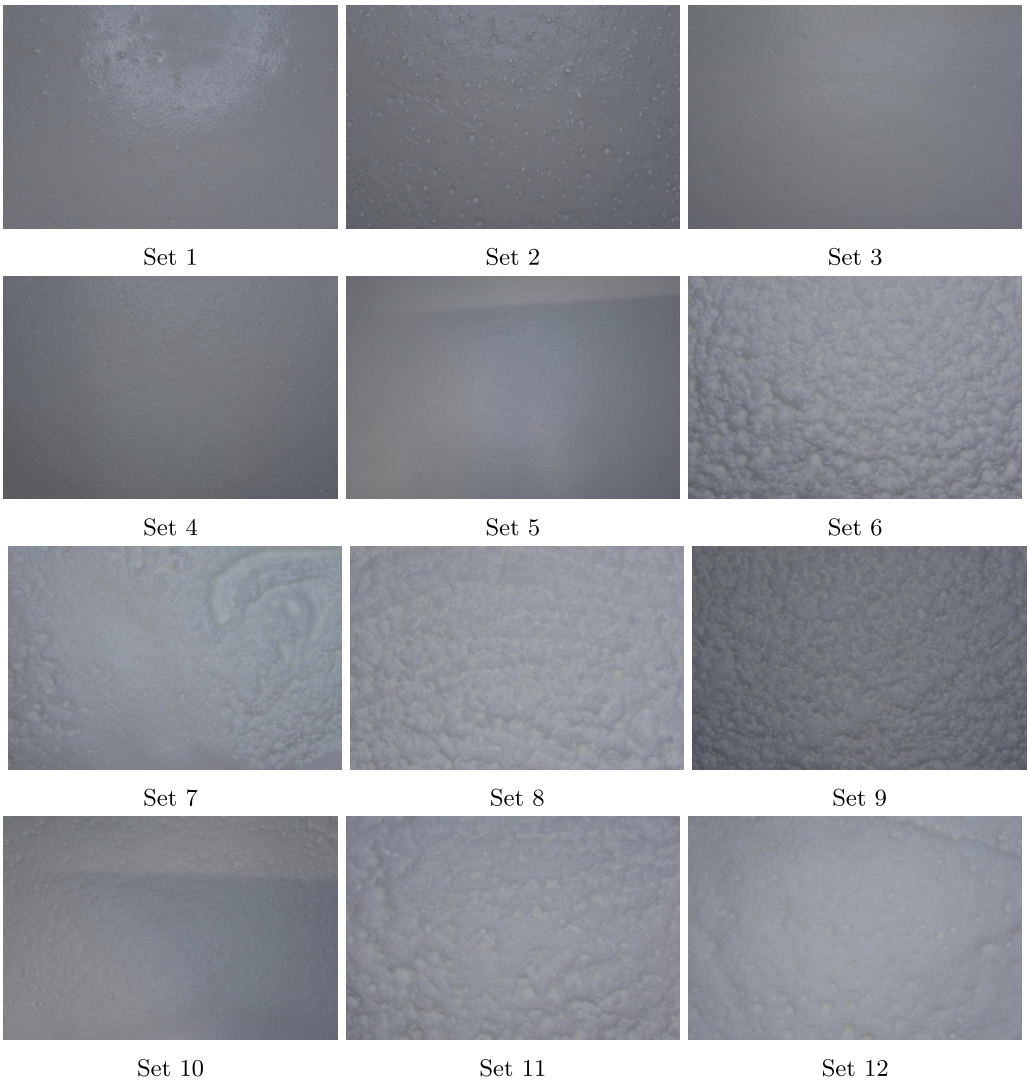


Fig. 2. The Target image of each set.

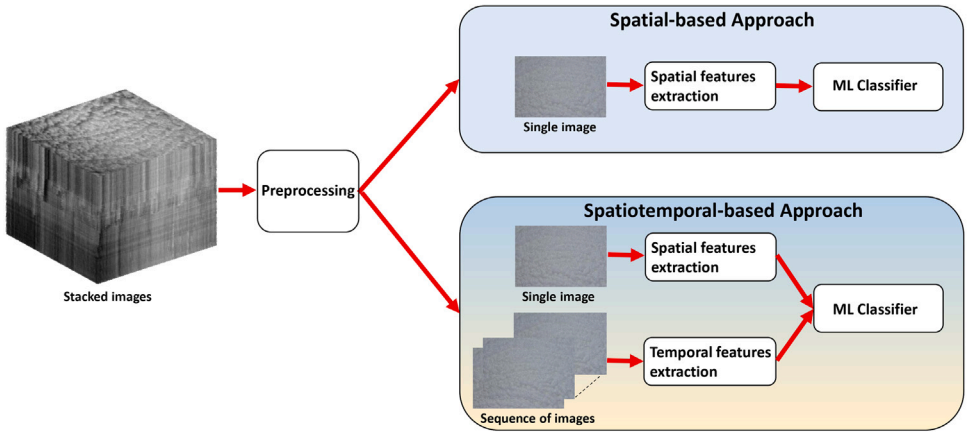
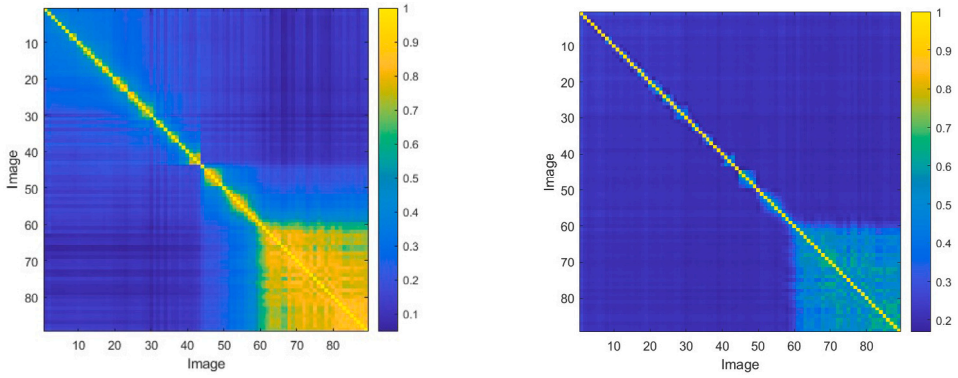


Fig. 3. Overview of the proposed framework for automatic estimation of coagulation time in cheese making. After a preprocessing step, the method includes a classification using spatial features extracted from the single images and a classification based on a combination of spatial and temporal information.

(a) Original image. (b) Image enhanced. (c) Image filtered.

Fig. 4. Sample image. Fig. 4a shows the original image, Fig. 4b the image enhanced by adaptive histogram equalization, and Fig. 4c the image after applying an additional filter based on local entropy.



(a) Autocorrelation matrix. (b) SSIM matrix.

Fig. 5. Autocorrelation matrix (Fig. 5a) and SSIM matrix (Fig. 5b) for a sample set. Minor changes are highlighted in yellow color, whereas significant ones are in blue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

even more the transformation process that occurs during the milk coagulation starting from the autocorrelation and SSIM matrices, we have considered the value changes by an average window and a timing window, as described below.

4.3.1. Identification of changes by average window

An average filtering has been applied to the autocorrelation matrix. The values along the diagonal of the filtered matrix are plotted as a function of time. The filter sizes are different: 5, 9, 15, and 21. The same filtering has been applied to the SSIM matrix. Fig. 6 reports the plot of the diagonal values at different filter sizes for the autocorrelation and the SSIM matrix. In both cases, the correlation values first decrease and then rapidly increase while approaching the CF-time. Afterward, these values show that a clear equilibrium has been reached. This phenomenon is very strong on the autocorrelation plot and less pronounced on the SSIM one. Moreover, the trend is much clearer if a filter size equal to 15 or 21 is chosen. Notably, in this example, the CF-time is equal to 61.

4.3.2. Identification of changes by timing window

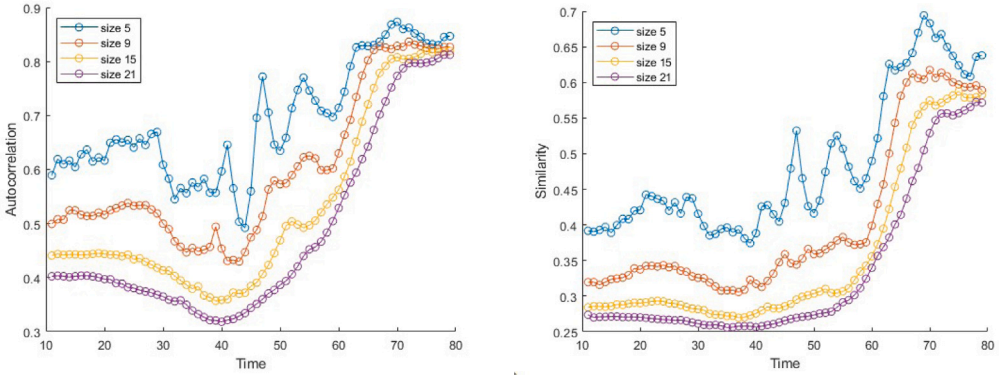
Another way to estimate the changes consists of evaluating the relation between an image and another one at a prefixed timing distance. So, both autocorrelation and SSIM can be helpful to understand to what extent images change over time. Starting from the first image, the autocorrelation between an image and another one at different time distances has been evaluated and plotted as a function of time. The same computation has been applied to the SSIM values. The distances have been chosen in accordance with the average filter size, so the values are 5, 10, 15, and 20. In Fig. 7, the trends of autocorrelation and SSIM evaluated at different timing windows are reported. Even in this case, when approaching the CF-time, i.e., 61, both the values start to rapidly increase, so correctly identifying the changes occurring during the coagulation process. Afterward, the values remain high and relatively constant. This is due to the chemical and physical modification that leads the milk to lose its characteristic of a liquid product and become gelatinous.

4.4. Spatiotemporal-based approach

Beyond the classical approach that consists of training models using features extracted from each image, as described in Section 4.2, a proper switching from prediction to classification typically consists of providing additional features that contain information about the sequence to be dealt with. The autocorrelation matrix and the SSIM matrices (illustrated in the previous section) are both helpful in detecting the information we are interested in and contain significant details for predicting the desired event.

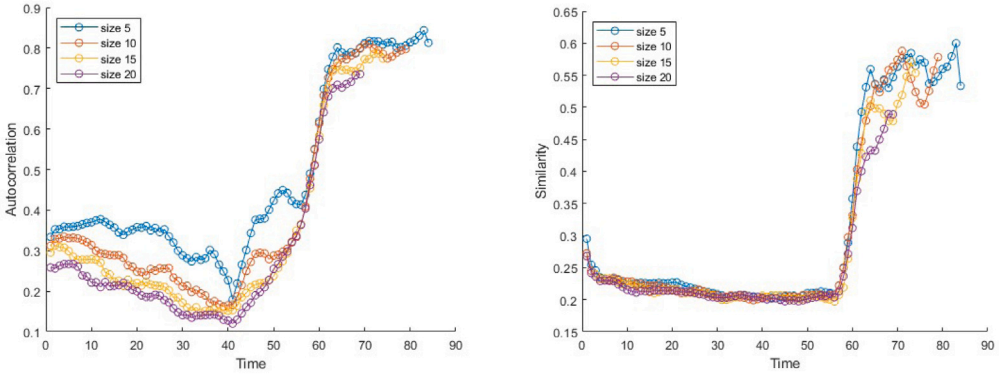
The moving window technique allows us to put this general strategy into practice. It consists of extracting relevant information from a fixed number of previous inputs with respect to the current one. In this proposal, the moving-window approach has been implemented by focusing on the matrix, called $\mathcal{M}$ hereinafter, generated by calculating, for each set, the autocorrelation (or similarity) among the given images after the preprocessing step. In particular, the generic element that occurs at position $\langle i, j \rangle$ of this 2D triangular matrix contains information about the correlation that holds between the $[i]$ image and the $[j]$ image, as described in Section 4.3 and visually showed in Fig. 5.

According to preliminary experiments, the moving window size (say *winsize*) has been set to 5. In fact, according to the results of this calibration stage, fewer elements did not guarantee the same performance, whereas more elements did not improve it. Windowing has been performed on top of $\mathcal{M}$ by looking at the correlation between the current image and the previous *winsize* images. In symbols, assuming that the current image occurs at the $[i]$ position, the values $\mathcal{M}[i, j]$, with $j = i-1, i-2, \ldots, i-winsize$ have been added to the features extracted from the image at hand. Beyond any formalization, this encoding accounts for the changes observed between the image in



(a) Autocorrelation trend. (b) SSIM trend.

Fig. 6. Trend of autocorrelation (Fig. 6a) and SSIM (Fig. 6b) values for average window of increasing size.



(a) Autocorrelation trend. (b) SSIM trend.

Fig. 7. Trend of autocorrelation (Fig. 7a) and SSIM (Fig. 7b) values for time window of increasing size.

question and the previous ones—up to the size of the mobile window. In conclusion, in the spatiotemporal approach, features extracted from images and features derived from the autocorrelation (or SSIM) matrix are concatenated to give information about spatial and time changes that occur in images during the coagulation process.

5. Experimental results and discussion

In this section, experimental results are presented. The experimental setup is described in Section 5.1. How autocorrelation and SSIM measures can be helpful in estimating coagulation time is discussed in Section 5.2. Sections 5.3 and 5.4 investigate the capability of creating proper models for dealing with the problem at hand. In particular, classifier models have been generated for each set and also on a set obtained by merging all sets. Proper summary tables follow hereinafter, reporting value and the standard deviation of each performance measure.

5.1. Experimental setup

In this research, all selected classifiers were trained with default parameters to prevent the generalization capability of the trained models from being influenced. Only the results obtained with the best shallow classifier (namely, an RF trained with 100 decision trees) have been reported for the sake of brevity. Regarding deep-learning-based classifiers, the underlying classification model is a standard multilayer perceptron (*MLP*), which has been trained using the progressive training technique. This technique belongs to the wide category of layer-wise training and is very suitable for developing *true MPLs*, i.e., models with multiple hidden layers, for it is not sensitive to the vanishing/exploding gradient problem (Armano, 2020; Glorot and Bengio, 2010).

In this research, various *MLP* architectures have been experimented with, using *MLP*s equipped with up to seven hidden layers. However, thanks to the adopted encoding technique, no improvement has been observed using more than four hidden layers. To adapt each *MLP* architecture to the selected encoding, two different choices have been made: (a) *MLP*s equipped with two hidden layers of 20 and 10 neurons, and (b) *MLP*s equipped with four hidden layers of 80, 40, 20, and 10 neurons. The former has been used with HC features, whereas the latter has DEEP features. The motivation for this choice is related to the number of features, the order of magnitude being 10^1 in the former case and 10^3 in the latter case. A detailed description of the hyperparameters used for the experimental evaluation is provided in Appendix, where Table A.13 shows the hyperparameters used to train the RF classifier, while Tables A.14 and A.15 present the hyperparameters chosen to train the two versions of MLP. In addition, we provide the code to conduct the experimental evaluation in a public GitHub repository[3]

The testing strategy was k-fold cross-validation, with $k = 10$. Relevant performance measures have been recorded for each experiment, in particular, F1, balanced accuracy, specificity, and sensitivity, as described in Section 3.3.

Experiments have been carried out ranging over five HC features and five features extracted by well-known CNNs. Tables 2 and 3 provide a short summary of these features.

[3] GitHub repository with the code realized for the experimental evaluation.

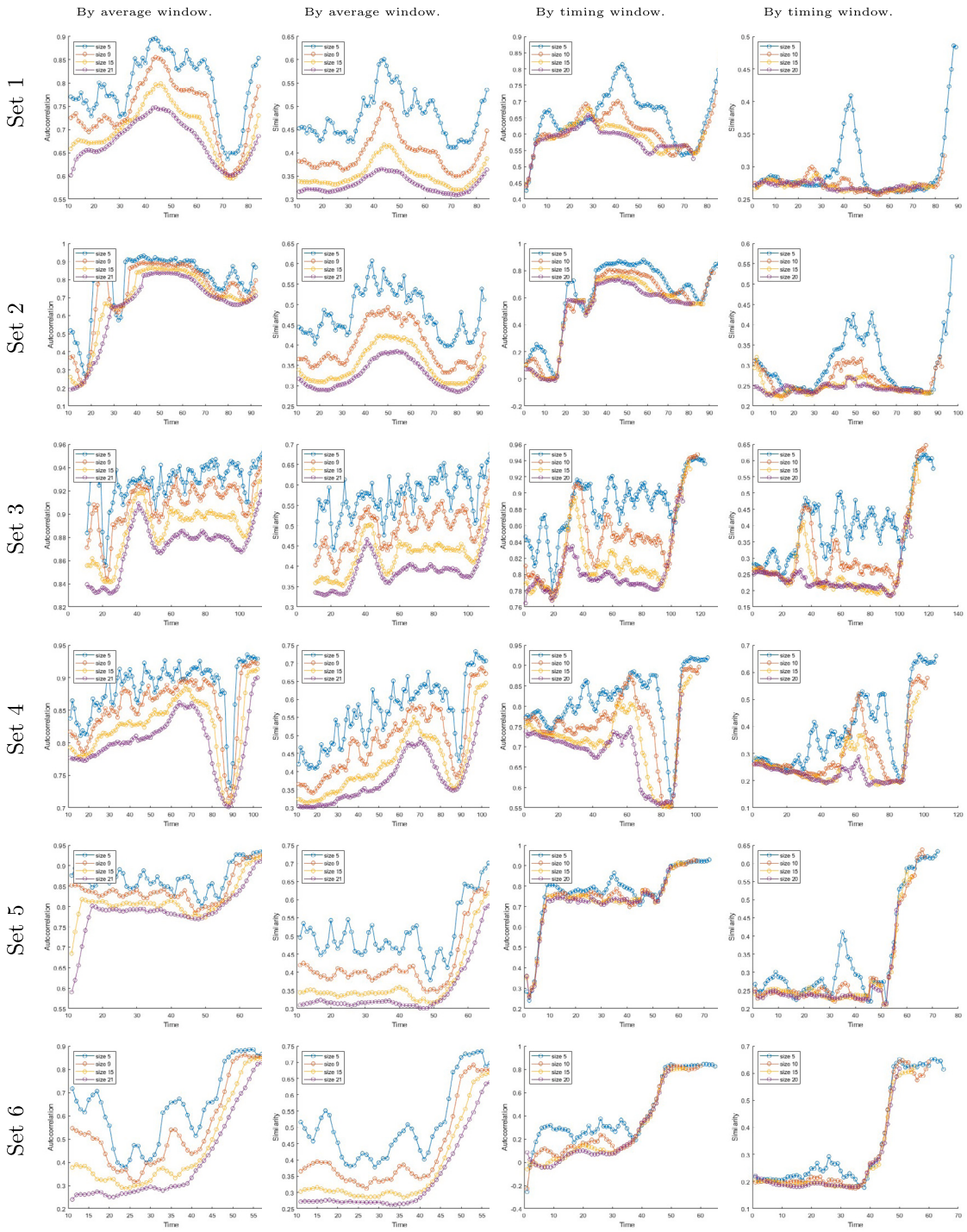


Fig. 8. Trend of autocorrelation and SSIM values by average window and by timing window for Sets 1 to 6.

5.2. Autocorrelation and SSIM for evaluating coagulation process

In Figs. 8 and 9, we show the autocorrelation and SSIM measures trends on all the sets of the analyzed dataset. Both the measures have been evaluated by the average window and by the timing window, as described in Section 4.3. All the graphs confirm how the milk coagulation process evolves. More precisely, after a rather chaotic and rapid start, an important event marks the beginning of the change in the milk state. In all image series, the graphs tend to get closer in the case of the average window approach and overlap in the case of the timing window when approaching the Target image that identifies this event. This confirms that both measures analyzed can detect informative changes.

It is worth pointing out in advance that experimental results highlight the existence of differences among the analyzed sets. The rationale behind this assertion lies in the multifaceted nature of coagulation, which is influenced not solely by the milk variety but also by additional variables, including rennet concentration and strength, temperature, milk preservation methods, as well as pH levels, and other chemical attributes of the milk, as highlighted in prior research (Stocco et al., 2021).

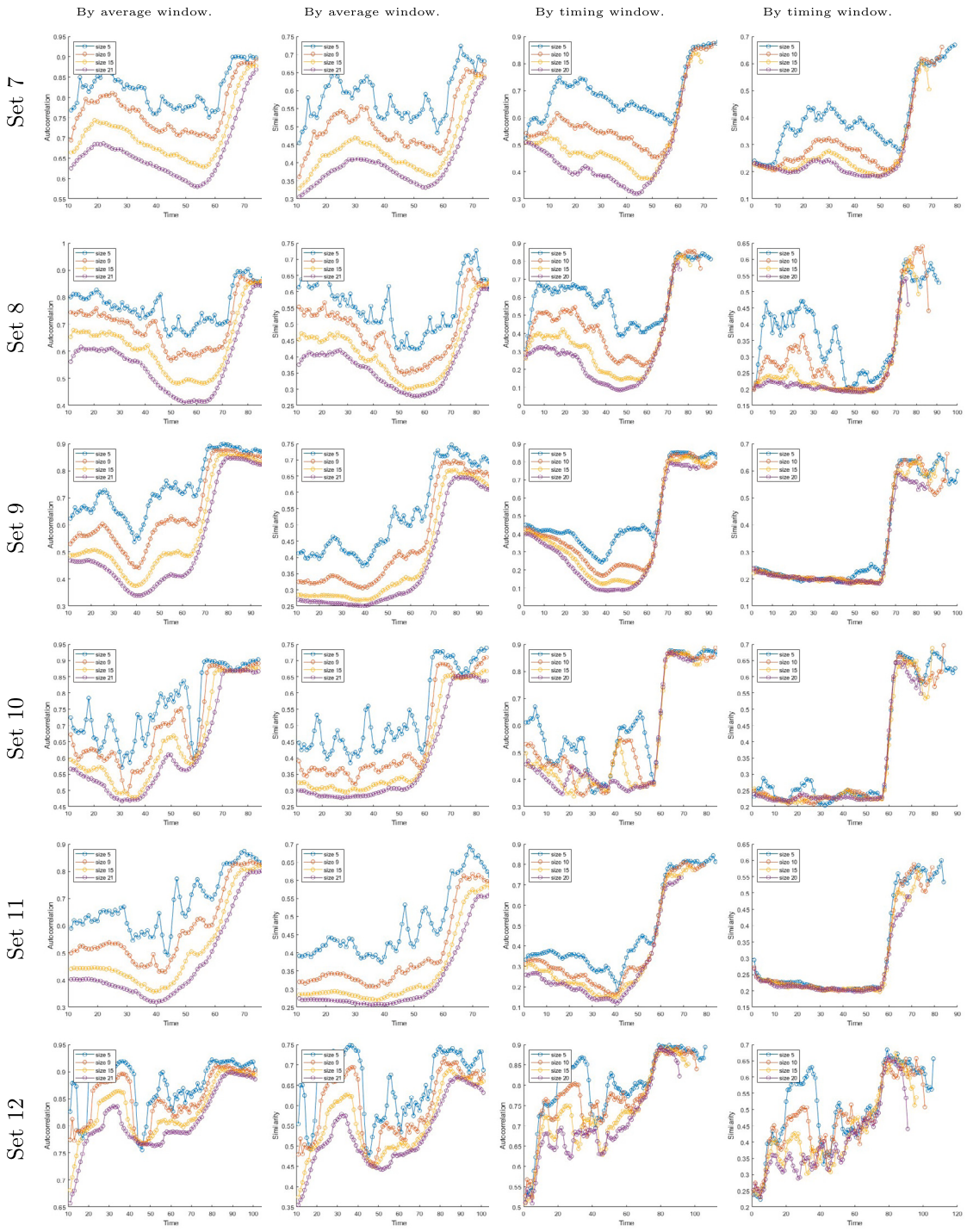


Fig. 9. Trend of autocorrelation and SSIM values by average window and by timing window for Sets 7 to 12.

However, beyond these particular conditions, the analysis based on the adopted features and techniques seems to contribute successfully to identifying the right time the curd begins, regardless of its development manner.

All these characteristics can influence the milk coagulation process, which strongly depends on environmental conditions. For example, the curd time is slow in Sets 1, 2, and 3, whereas it is fast in Sets 6, 10, and 11. In addition, the curd is slow in Set 12 but forms an inconsistent clot. However, beyond these particular conditions, the analysis based on the adopted features and techniques contributes successfully to the identification of the right time in which the curd begins, regardless of its development manner.

5.3. Experiments carried out on single sets

The first benchmarking experiments have been performed by training separate models on each selected set. As pointed out, experimental results have been averaged downstream of a k-fold cross-validation

Table 5
Results obtained by training a *Random Forest* classifier on each set by the spatial-based approach with HC features. Average performance and standard deviation (the latter within round brackets) are reported for each feature group.

Group		Set	F1	Accuracy	Specificity	Sensitivity
CH_5	*best*	7	1.00 (0.00)	0.99 (0.01)	1.00 (0.00)	0.99 (0.01)
	worst	8	0.44 (0.03)	0.53 (0.03)	0.52 (0.03)	0.53 (0.03)
	avg	*all*	0.75 (0.17)	0.80 (0.17)	0.82 (0.13)	0.77 (0.16)
CH2_5	*best*	8	0.95 (0.02)	0.95 (0.03)	0.96 (0.03)	0.93 (0.02)
	worst	7	0.41 (0.02)	0.44 (0.02)	0.36 (0.05)	0.52 (0.03)
	avg	*all*	0.74 (0.16)	0.77 (0.15)	0.76 (0.17)	0.77 (0.13)
HARri	*best*	11	0.91 (0.03)	0.92 (0.04)	0.90 (0.03)	0.93 (0.04)
	worst	8	0.47 (0.03)	0.59 (0.03)	0.63 (0.03)	0.54 (0.04)
	avg	*all*	0.65 (0.19)	0.71 (0.19)	0.72 (0.13)	0.69 (0.17)
LBP_18	*best*	9	0.81 (0.09)	0.84 (0.07)	0.80 (0.08)	0.87 (0.03)
	worst	8	0.43 (0.02)	0.47 (0.02)	0.45 (0.03)	0.48 (0.02)
	avg	*all*	0.58 (0.13)	0.64 (0.12)	0.64 (0.14)	0.63 (0.14)
LM_5	*best*	7	0.99 (0.01)	0.99 (0.02)	0.99 (0.01)	0.99 (0.01)
	worst	12	0.39 (0.02)	0.46 (0.02)	0.41 (0.03)	0.51 (0.03)
	avg	*all*	0.74 (0.17)	0.79 (0.16)	0.80 (0.13)	0.77 (0.15)

Table 6
Results obtained by training a *Random Forest* classifier on each set by the spatial-based approach with deep features. Average performance and standard deviation (the latter within round brackets) are reported for each feature group.

Group		Set	F1	Accuracy	Specificity	Sensitivity
DarkNet-53	*best*	7	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)	0.98 (0.02)
	worst	8	0.44 (0.02)	0.58 (0.02)	0.62 (0.02)	0.53 (0.03)
	avg	*all*	0.82 (0.14)	0.85 (0.13)	0.86 (0.10)	0.83 (0.11)
GoogLeNet	*best*	9	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)
	worst	3	0.42 (0.02)	0.56 (0.02)	0.59 (0.08)	0.53 (0.03)
	avg	*all*	0.80 (0.18)	0.85 (0.17)	0.86 (0.12)	0.83 (0.15)
Inception-v3	*best*	8	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)
	worst	4	0.62 (0.05)	0.66 (0.09)	0.65 (0.03)	0.67 (0.04)
	avg	*all*	0.87 (0.10)	0.89 (0.10)	0.89 (0.09)	0.88 (0.08)
ResNet-18	*best*	7	1.00 (0.00)	0.99 (0.01)	1.00 (0.00)	0.99 (0.00)
	worst	8	0.46 (0.02)	0.63 (0.02)	0.72 (0.03)	0.54 (0.02)
	avg	*all*	0.71 (0.15)	0.77 (0.15)	0.81 (0.11)	0.73 (0.14)
ResNet-101	*best*	7	1.00 (0.00)	0.99 (0.01)	1.00 (0.00)	0.99 (0.00)
	worst	12	0.46 (0.02)	0.59 (0.02)	0.65 (0.01)	0.53 (0.00)
	avg	*all*	0.85 (0.15)	0.88 (0.15)	0.89 (0.11)	0.86 (0.13)

Table 7
Results obtained by training *MLP* models on each set by the spatial approach with HC features. Average performance and standard deviation (the latter within round brackets) are reported for each feature group.

Group		Set	F1	Accuracy	Specificity	Sensitivity
CH_5	*best*	9	0.97 (0.04)	0.99 (0.01)	0.97 (0.04)	1.00 (0.00)
	worst	2	0.00 (0.00)	0.50 (0.03)	1.00 (0.00)	0.00 (0.00)
	avg	*all*	0.70 (0.12)	0.85 (0.04)	0.95 (0.03)	0.75 (0.10)
CH2_5	*best*	9	0.99 (0.02)	0.99 (0.01)	0.99 (0.02)	1.00 (0.00)
	worst	2	0.00 (0.00)	0.50 (0.04)	1.00 (0.00)	0.00 (0.00)
	avg	*all*	0.60 (0.15)	0.80 (0.05)	0.95 (0.04)	0.64 (0.16)
HARri	*best*	9	0.81 (0.09)	0.86 (0.07)	0.86 (0.09)	0.86 (0.12)
	worst	3	0.04 (0.07)	0.51 (0.09)	0.97 (0.07)	0.05 (0.11)
	avg	*all*	0.46 (0.15)	0.69 (0.15)	0.90 (0.07)	0.48 (0.22)
LBP_18	*best*	9	0.88 (0.05)	0.90 (0.03)	0.94 (0.05)	0.86 (0.08)
	worst	2	0.08 (0.24)	0.54 (0.03)	1.00 (0.01)	0.07 (0.20)
	avg	*all*	0.71 (0.12)	0.83 (0.11)	0.92 (0.05)	0.73 (0.16)
LM_5	*best*	9	1.00 (0.01)	1.00 (0.01)	1.00 (0.01)	1.00 (0.00)
	worst	2	0.00 (0.00)	0.50 (0.05)	1.00 (0.00)	0.00 (0.00)
	avg	*all*	0.71 (0.13)	0.86 (0.04)	0.96 (0.04)	0.76 (0.14)

procedure, with $k = 5$. To avoid running into too many details (in fact, $5 + 5$ groups of features applied to 12 sets would generate a total of 120 rows), for each feature group, the worst and the best results are reported, together with information about the average behavior.

Table 8
Results obtained by training *MLP* models on each set by the spatial approach, with features extracted by means of well-known CNNs. Average performance and standard deviation (the latter within round brackets) are reported for each feature group.

Group		Set	F1	Accuracy	Specificity	Sensitivity
DarkNet-53	*best*	11	0.98 (0.02)	0.99 (0.01)	0.98 (0.02)	1.00 (0.00)
	worst	2	0.74 (0.26)	0.90 (0.02)	0.96 (0.03)	0.84 (0.29)
	avg	*all*	0.89 (0.08)	0.95 (0.03)	0.95 (0.04)	0.93 (0.08)
GoogLeNet	*best*	8	0.98 (0.04)	0.99 (0.03)	0.98 (0.05)	1.00 (0.00)
	worst	2	0.48 (0.34)	0.77 (0.05)	0.95 (0.03)	0.59 (0.42)
	avg	*all*	0.87 (0.08)	0.93 (0.04)	0.94 (0.05)	0.92 (0.09)
Inception-v3	*best*	8	0.96 (0.05)	0.99 (0.03)	0.98 (0.03)	1.00 (0.00)
	worst	3	0.80 (0.08)	0.91 (0.02)	0.96 (0.03)	0.86 (0.15)
	avg	*all*	0.92 (0.06)	0.96 (0.03)	0.95 (0.04)	0.96 (0.05)
ResNet-18	*best*	9	0.97 (0.04)	0.98 (0.04)	0.96 (0.06)	1.00 (0.00)
	worst	5	0.85 (0.08)	0.92 (0.07)	0.90 (0.11)	0.93 (0.09)
	avg	*all*	0.93 (0.06)	0.98 (0.03)	0.96 (0.04)	0.99 (0.03)
ResNet-101	*best*	7	0.99 (0.03)	1.00 (0.01)	0.99 (0.02)	1.00 (0.00)
	worst	5	0.80 (0.09)	0.88 (0.06)	0.88 (0.09)	0.87 (0.10)
	avg	*all*	0.93 (0.05)	0.98 (0.03)	0.96 (0.04)	0.99 (0.02)

5.3.1. Spatial-based approach

Here, we present the results obtained with the two classification methodologies adopted in the context of the spatial-based approach. More specifically, Tables 5 and 6 report the results obtained with the best shallow learning classifier (viz. *Random Forest*) trained with HC and DEEP features, respectively. In addition, Tables 7 and 8, report the performance measures obtained with the deep learning classifier (viz. *MLP*) trained with HC and deep spatial-based features, respectively. Each table provides a detailed breakdown of performance measures, including F1, accuracy, specificity, and sensitivity, for different feature groups across the twelve sets.

Spatial-based approach with a shallow classifier. In Table 5, the results for handcrafted features demonstrate distinct performance variations among feature groups. The best performance in terms of F1 was achieved by the CH_5 feature group on Set 7, with a score of 1.00. Conversely, the worst performance with the same feature group was obtained on Set 8, achieving an F1 of 0.44. On average, the performance across all sets for the HC features ranged from 0.58 to 0.75 for the F1. Similarly, the *CH2_5 HARri*, *LBP_18*, and *LM_5* groups exhibit varying performances across different sets. Notably, the best performances in these groups are achieved in Sets 8, 11, 9, and 7, respectively. Conversely, the worst performances are observed in Sets 7, 8, and 12 for these groups.

Moving to Table 6, we observe that certain DEEP features, such as *DarkNet-53* and *ResNet-101*, consistently yield high performance across various sets. For example, *DarkNet-53* achieves a 0.99 F1 on set 7, and both ResNets obtained a 1.00 F1, highlighting its robust feature representation. On the other hand, the *GoogLeNet* group demonstrates variability, with set 3 showing lower performance compared to other sets, with an F1 of 0.42. On average, the performance across all sets for the DEEP features ranged from 0.71 to 0.87 for the F1, indicating moderate to good classification performance.

Overall, both tables demonstrate that the performance of the *Random Forest* classifier varies across different sets and feature groups. DEEP features generally tend to outperform handcrafted features, as seen in the higher average F1 values. However, there is still variability in performance, emphasizing the challenging task in terms of classification approaches.

Spatial-based approach with a deep classifier. Moving to the deep classifier, the results obtained by training MLP models on each set using the spatial approach with HC and DEEP features are reported in Tables 7 and 8, respectively. In Table 7, the best performance in terms of F1 was achieved using the *CH_5* features on Set 9, with a score of 0.97. Conversely, the worst performance was observed with the same feature

Table 9
Results obtained by training *MLP* models on each set by the spatiotemporal approach, i.e., using HC features combined with features extracted from autocorrelation matrix through a mobile window technique. Average performance and standard deviation (the latter within round brackets) are reported for each feature group.

Group		Set	F1	Accuracy	Specificity	Sensitivity
CH_5	*best*	11	0.98 (0.02)	0.99 (0.01)	0.99 (0.02)	0.98 (0.03)
	worst	2	0.00 (0.00)	0.50 (0.03)	1.00 (0.00)	0.00 (0.00)
	avg	*all*	0.83 (0.08)	0.91 (0.03)	0.97 (0.03)	0.84 (0.09)
CH2_5	*best*	9	0.99 (0.01)	0.99 (0.01)	0.99 (0.02)	0.99 (0.02)
	worst	2	0.00 (0.00)	0.50 (0.00)	1.00 (0.00)	0.00 (0.00)
	avg	*all*	0.80 (0.10)	0.89 (0.04)	0.97 (0.03)	0.80 (0.13)
HARri	*best*	9	0.98 (0.03)	0.98 (0.02)	1.00 (0.00)	0.95 (0.06)
	worst	2	0.03 (0.10)	0.55 (0.04)	0.99 (0.03)	0.10 (0.30)
	avg	*all*	0.71 (0.12)	0.84 (0.04)	0.96 (0.04)	0.71 (0.15)
LBP_18	*best*	9	0.98 (0.02)	0.98 (0.01)	0.99 (0.02)	0.97 (0.03)
	worst	2	0.43 (0.29)	0.68 (0.02)	0.98 (0.02)	0.38 (0.28)
	avg	*all*	0.86 (0.07)	0.92 (0.03)	0.97 (0.03)	0.86 (0.11)
LM_5	*best*	9	0.98 (0.02)	0.99 (0.01)	0.99 (0.02)	0.99 (0.02)
	worst	2	0.00 (0.00)	0.50 (0.04)	1.00 (0.00)	0.00 (0.00)
	avg	*all*	0.83 (0.09)	0.91 (0.04)	0.98 (0.04)	0.84 (0.09)

group on Set 2, achieving an F1 of 0.00. On average, the performance across all sets for the HC features ranged from 0.46 to 0.71 for the F1, indicating moderate classification performance.

Table 8 shows how *DarkNet-53* feature group exhibited the best performance in terms of F1 on Set 11, achieving a score of 0.98. Conversely, the worst performance was observed with the *GoogLeNet* feature group on Set 2, with an F1 of 0.48. On average, the performance across all sets for the DEEP features ranged from 0.87 to 0.93 for the F1, indicating comparatively better classification performance than the HC features.

Overall, the results suggest that classifier models trained using DEEP features tend to outperform those trained with HC features on the different sets using the spatial-based approach. Additionally, there is variability in performance across different sets, highlighting the more pronounced difficulties in achieving optimal classification results in all sets.

5.3.2. Spatiotemporal-based approach

The outcomes attained using the *MLP* approach, employing the merging of individual image features with temporal data given by the autocorrelation, are reported hereinafter. This consideration stems from the remarkable outcomes observed on *MLP* in the spatial approach delineated in Section 5.3.1. For brevity's sake, we present the findings derived from integrating autocorrelation, as they exhibit nearly identical behavior to the SSIM.

Quantitative results. The results obtained across different encodings are reported in Tables 9 and 10.

Table 9 highlights that the *CH_5* feature group achieved the best performance on Set 11, with an F1 of 0.98, indicating high accuracy and balanced precision–recall trade-off. Conversely, the worst performance was observed with the same feature group on Set 2, achieving an F1 of 0.00, indicating poor classification capability. On average, the performance across all sets for the combined HC and autocorrelation matrix features ranged from 0.71 to 0.86 for the F1, showcasing moderate classification performance with some variability.

As shown in Table 10, the *DarkNet-53* feature group exhibited the best performance on Set 11, with an F1 of 0.98, demonstrating robust classification capability. Conversely, the worst performance was observed with the *GoogLeNet* feature group on Set 2, achieving an F1 of 0.51. On average, the performance across all sets for the combined DEEP features and autocorrelation matrix features ranged from 0.87 to 0.93 for the F1, indicating good classification performance with relatively low variability.

All experiments confirm that using single image features in combination with sequence information allows excellent results in terms of performance to be obtained. A comparison between Table 9 and Table 10 points out that the latter performs better and highlights the effectiveness of combining spatiotemporal features extracted using different techniques for enhancing the performance. The values of standard deviation recorded with varying test runs grant results statistical significance. Similar results have been achieved using SSIM instead of the autocorrelation matrix to detect temporal information; in other words, both measures can be used to derive sequence information.

Qualitative results. We conducted a further analysis to demonstrate how features extracted from autocorrelation matrices have contributed to enhancing the results achieved solely using features, HC, or DEEP. To highlight this aspect effectively, we specifically chose the best and worst HC features, as well as the best and worst DEEP features, based on their average F1-score reported in Tables 7–10. Our selection identified *LBP_18* as the best HC feature (0.71 F1-score in the spatial approach vs. 0.86 in the spatiotemporal approach) and *HARri* as the worst HC feature (0.46 vs. 0.71), while *ResNet-101* emerged as the best deep feature (0.93 vs. 0.93) and *GoogLeNet* as the worst deep feature (0.87 vs. 0.94).

Fig. 10 illustrates the behavior of each of the four selected features under two scenarios: when used independently (spatial approach, depicted by the blue line) and when integrated with the autocorrelation features (spatiotemporal approach, indicated by the red line). Furthermore, we compare these behaviors against the average F1 of their respective categories, either in terms of the spatial approach (represented by the yellow line) or the spatiotemporal approach (illustrated by the green line).

Delving into the specifics of the trends, it becomes evident that each feature experienced an enhancement through the combination. Notably, both of the considered HC features exhibited significant improvements, as highlighted in Figs. 10(a) and 10(c). For instance, examining the performance in Set 3, the F1-score for *LBP_18* increased from 0.73 to 0.86, while *HARri* improved from 0.04 to 0.51. This observation underscores the utility of autocorrelation features in improving the performance of HC features.

Conversely, the enhancements related to the DEEP features are less pronounced, although in the majority of sets, both *ResNet-101* and *GoogLeNet* demonstrated improvements.

Fig. 10 further elucidates the inherent challenges present in certain subsets, which were previously highlighted in Tables 5–10, with specific reference to the scores achieved by the HC features. For instance, all features exhibited subpar classification results in Set 2 (although *ResNet-101* features achieved an F1-score of 0.75), while *HARri* encountered difficulties across Sets 1 to 8 with the spatial approach. Furthermore, from a broader perspective, Fig. 10 underscores the clear superiority of DEEP features over HC features in both settings, coupled with greater robustness.

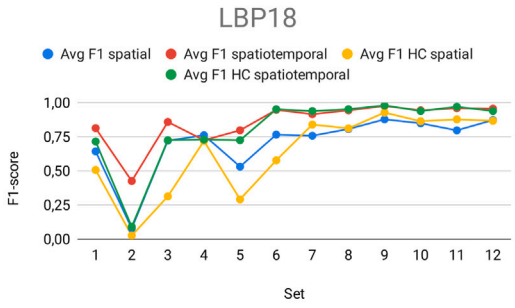
5.4. Experiments carried after merging all sets

The second benchmarking experiments have been performed by training MLP models on a *merge* of the selected sets, using the spatiotemporal approach, which combines both HC (see Table 11) and DEEP features (see Table 10) with features extracted from autocorrelation matrices using a mobile window technique. This decision stems from the remarkable outcomes observed with the spatiotemporal approach delineated in Section 5.3.2. Again, experimental results have been averaged downstream of a k-fold cross-validation procedure. For the sake of brevity, we solely present the findings derived from integrating autocorrelation, as they exhibit nearly identical behavior to the SSIM.

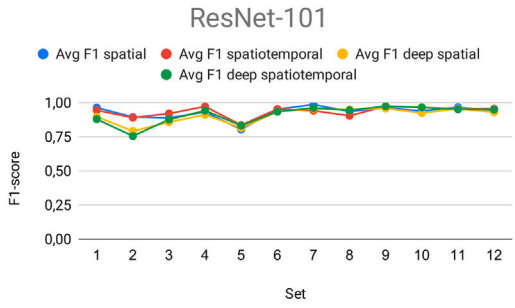
In particular, Table 11 shows all feature groups achieved high performance across all measures, with F1 values ranging from 0.98

Table 10
Results obtained by training *MLP* models on each set by the spatiotemporal approach, i.e., using features extracted by means of well-known deep neural networks together with features extracted from autocorrelation matrix through a mobile window technique. Average performance and standard deviation (the latter within round brackets) are reported for each feature group.

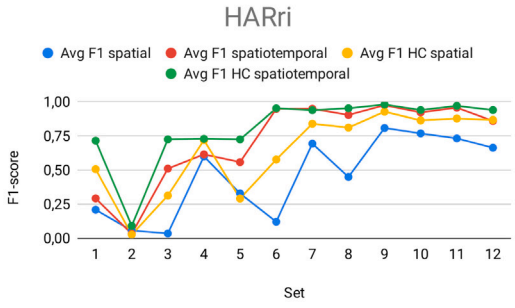
Group		Set	F1	Accuracy	Specificity	Sensitivity
DarkNet-53	*best*	11	0.98 (0.02)	0.99 (0.01)	0.98 (0.02)	1.00 (0.00)
	worst	2	0.82 (0.17)	0.94 (0.04)	0.96 (0.03)	0.92 (0.20)
	avg	*all*	0.92 (0.06)	0.96 (0.03)	0.95 (0.04)	0.96 (0.05)
GoogLeNet	*best*	7	0.99 (0.03)	1.00 (0.01)	1.00 (0.01)	1.00 (0.00)
	worst	2	0.51 (0.30)	0.79 (0.07)	0.92 (0.08)	0.65 (0.34)
	avg	*all*	0.87 (0.08)	0.94 (0.03)	0.95 (0.04)	0.93 (0.08)
Inception-v3	*best*	12	0.98 (0.02)	0.99 (0.01)	0.98 (0.02)	1.00 (0.00)
	worst	2	0.78 (0.27)	0.94 (0.06)	0.97 (0.02)	0.90 (0.30)
	avg	*all*	0.93 (0.06)	0.97 (0.03)	0.96 (0.04)	0.98 (0.05)
ResNet-18	*best*	9	0.99 (0.02)	0.99 (0.02)	0.98 (0.03)	1.00 (0.00)
	worst	2	0.77 (0.13)	0.94 (0.03)	0.96 (0.03)	0.92 (0.17)
	avg	*all*	0.92 (0.06)	0.97 (0.03)	0.96 (0.04)	0.98 (0.04)
ResNet-101	*best*	9	0.97 (0.02)	0.99 (0.02)	0.97 (0.03)	1.00 (0.00)
	worst	5	0.84 (0.07)	0.91 (0.05)	0.89 (0.07)	0.92 (0.09)
	avg	*all*	0.93 (0.05)	0.98 (0.03)	0.96 (0.04)	0.99 (0.01)



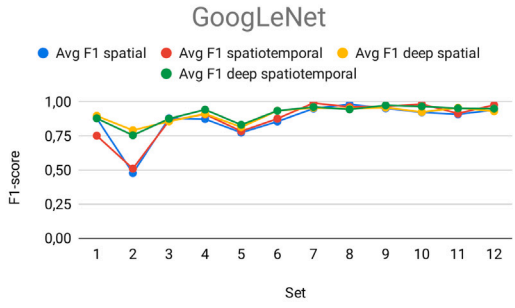
(a) Performance of the best HC features with spatial and spatiotemporal approaches, and comparison against the average HC category across the twelve sets included in the dataset.



(b) Performance of the best deep features with spatial and spatiotemporal approaches, and comparison against the average deep category across the twelve sets included in the dataset.



(c) Performance of the least effective HC features with spatial and spatiotemporal approaches, and comparison against the average HC category across the twelve sets included in the dataset.



(d) Performance of the least effective deep features with spatial and spatiotemporal approaches, and comparison against the average deep category across the twelve sets included in the dataset.

Fig. 10. Qualitative and comparative analysis of extracted features: this figure examines four distinct features. Fig. 10a and 10b present the best HC and deep features, respectively. Conversely, Fig. 10a and 10b illustrates the least effective HC and deep features, respectively. The evaluation assesses their performance using both spatial and spatiotemporal approaches, comparing them against the average within their respective categories across the twelve sets included in the dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

(*LBP_18*) to 0.99. These results indicate that combining HC features with features extracted from autocorrelation matrices leads to an improved and robust classification performance with high accuracy, specificity, and sensitivity. Only *LBP_18* struggles in terms of specificity, with a score of 0.92.

Similarly to what was observed in Table 11, Table 12 reports that the DEEP features consistently achieved high performance across all measures. Specifically, all feature groups achieved an F1 of 0.99, indicating excellent classification capability. Overall, these results demonstrate that combining features extracted from CNNs with those from

Table 11
Results obtained by training *MLP* models on a set obtained merging all available data by the spatiotemporal approach, i.e., using HC features together with features extracted from $\mathcal{M}$ through a mobile window technique. Average performance and standard deviation (the latter within round brackets) are reported for each feature group.

Group	F1	Accuracy	Specificity	Sensitivity
CH_5	0.99 (0.00)	0.98 (0.00)	0.97 (0.03)	0.99 (0.00)
CH2_5	0.99 (0.00)	0.99 (0.00)	0.98 (0.02)	0.99 (0.00)
HARri	0.99 (0.00)	0.98 (0.00)	0.97 (0.03)	0.99 (0.01)
LPB18	0.98 (0.00)	0.95 (0.01)	0.92 (0.08)	0.98 (0.01)
LM_5	0.99 (0.00)	0.99 (0.00)	0.98 (0.03)	0.99 (0.00)

Table 12
Results obtained by training *MLP* models on a set obtained merging all available data by the spatiotemporal approach, i.e., using features extracted from CNNs together with features extracted from $\mathcal{M}$ through a mobile window technique. Average performance and standard deviation (the latter within round brackets) are reported for each feature group.

Group	F1	Accuracy	Specificity	Sensitivity
DarkNet-53	0.99 (0.00)	0.97 (0.01)	0.95 (0.07)	0.99 (0.01)
GoogLeNet	0.99 (0.00)	0.99 (0.00)	0.99 (0.02)	0.99 (0.00)
Inception-v3	0.99 (0.00)	0.97 (0.00)	0.95 (0.01)	0.99 (0.00)
ResNet-18	0.99 (0.00)	0.98 (0.01)	0.97 (0.04)	0.99 (0.00)
ResNet-101	0.99 (0.00)	0.98 (0.01)	0.96 (0.04)	0.99 (0.01)

autocorrelation matrices enhances the classification performance of MLP models, achieving high accuracy and robustness.

Significantly, both the HC and DEEP features demonstrate marginally higher sensitivity in contrast to specificity, suggesting an enhanced capability to accurately classify positive instances.

Overall, all conducted experiments affirm that leveraging single image features in conjunction with sequence information yields very good results in terms of performance, even when utilizing a single set derived from the amalgamation of selected sets. A comparison between Tables 11 and 12 does not provide substantial evidence regarding the superiority of HC vs. DEEP features, or vice versa, when considering the amalgamation of all available sets. Furthermore, the statistical significance of the results is upheld by the observed standard deviation values across varying test runs.

6. Discussion

This section provides a comprehensive review of the obtained results (refer to Section 6.1), explores the potential implications within the cheese industry (refer to Section 6.2), and presents a comparison with the previous research (see Section 6.3).

6.1. On the performance results

The experiments conducted in this study provide valuable insights into the performance of different classification methodologies when applied to single sets as well as merged sets.

Firstly, the results indicate significant variability in performance across different feature groups and classifiers when considering single sets. Although handcrafted features exhibit moderate classification performance, they demonstrate considerable variability across sets. In contrast, DEEP features consistently outperform HC features on average, showcasing better classification capabilities. However, both approaches encounter difficulties in achieving optimal performance across all sets, emphasizing the complexity of the classification task.

Secondly, integrating sequence information with single image features through the spatiotemporal-based approach yields notable improvements in classification performance. Combining HC or DEEP features with features extracted from autocorrelation matrices results in enhanced classification accuracy and robustness. DEEP features, in particular, demonstrate excellent classification capability when integrated with sequence information, achieving high accuracy and robustness across all measures.

Furthermore, the qualitative analysis highlights the effectiveness of autocorrelation features in enhancing the performance of both HC and DEEP features. Although the improvements are more pronounced for HC features, DEEP features also benefit from integrating sequence information.

Finally, we have also considered the use of end-to-end deep learning architectures. However, we have determined that the application of Vision Transformer (ViT), as well as fine-tuning CNNs, was not suitable for the specific focus of our study.

Specifically, ViT leverages self-attention mechanisms to analyze image patches and emphasize relevant information, which is effective for tasks requiring a comprehensive understanding of static images. However, our study involves sequential image data capturing the evolving stages of cheese ripening, necessitating a model capable of capturing temporal correlations across these sequences (Dosovitskiy et al., 2021; Han et al., 2022).

Our consideration of ViT initially centered on utilizing its feature extraction capabilities for comparative analysis against models leveraging HC and DEEP features. However, the direct comparison was constrained by our use of CNNs trained on a limited Imagenet1k dataset, precluding a comprehensive evaluation against ViT's performance measures. Notably, ViT has demonstrated exceptional efficacy when pre-trained at scale and then adapted to tasks with fewer data points (Dosovitskiy et al., 2021).

Furthermore, the constraints posed by our dataset's size and class imbalance made training a ViT model from scratch unfeasible, paralleling the challenges encountered with CNNs. Our dataset was tailored for a preliminary investigation into the viability of a vision-based system for milk coagulation time assessment in cheese production—a pilot initiative aimed at laying foundational groundwork rather than comprehensive model optimization.

6.2. On the implications in cheese industry

This research focuses on automating the detection of curd-firming time during cheese production by utilizing CV and ML techniques. The theoretical implications of this work lie in the advancement of automated, artificial intelligence-based methods for enhancing production efficiency and ensuring consistent product quality in the cheese-making process. By leveraging image features and temporal relationships among images, the proposed approach offers a novel way to monitor the curd formation process in real time, providing insights into curd firmness and predicting optimal cutting times.

From a practical standpoint, this research makes several key contributions to the field. First, it introduces CM-IDB, the first publicly available image dataset related to the cheese-making process. Second, it presents an innovative approach for detecting curd-firming time, addressing a critical challenge in cheese production, as precise cutting time significantly influences the quality and quantity of the final cheese product (Arango and Castillo, 2018; Gao et al., 2022; Guinee, 2021). Third, the research offers a hybrid solution that combines image features and temporal characteristics to enhance the accuracy of curd-firming time detection.

This research introduces a unique and practical approach by emphasizing the crucial determination of curd-firming time and utilizing a streamlined yet effective technical setup. The proposed system's simplicity and accessibility hold promise for enhancing process control and optimizing cheese production, ultimately leading to improved efficiency and product quality. Additionally, the incorporation of image features and temporal relationships distinguishes this research by establishing a solid foundation for identifying curd-firming time within the CM-IDB dataset.

6.3. Comparison with previous research

The results of the study indicate that the proposed CV and ML-based approach significantly improve the accuracy of curd-firming time detection during cheese production.

Compared with existing research, our study employs a combination of CV and ML techniques, which is unexplored in the task of cutting time detection, as well as the approach is innovative from the point of view of the input used, based on images acquired with an ordinary photo camera. In addition, the proposed approaches leverages sequence information from images, differently from the traditional approaches employed in this context, which primarily focus on task-specific techniques, such as electrical, thermal, optical, and ultrasonic methods (Gao et al., 2022; Feng et al., 2021; Vacca et al., 2020; Lazouskaya et al., 2021; Hwang et al., 2022; Hass et al., 2015; Tabayehnejad et al., 2012; Budelli et al., 2017). This research distinguishes itself by targeting the pivotal stage of cheese production—the determination of curd-firming time with a combination of CV and ML techniques. This stage has not been previously explored in the literature with this kind of approach and setup.

More precisely, the proposed approach utilizes a straightforward setup consisting of a camera connected to a computer, avoiding more complex configurations. The key advantage of this setup is its non-invasive and non-destructive nature, allowing for the monitoring of the cheese production process without interfering with or damaging the product. Additionally, the simplicity of the setup facilitates easy implementation, making it an accessible solution for practical applications in the cheese industry.

We also released CM-IDB, the first publicly available image dataset related to cheese-making. This dataset enhances reproducibility and provides a benchmark for future studies. In this context, previous studies, as indicated in Section 2, used private datasets of different types that were not disclosed, limiting the ability to reproduce and validate their findings.

Finally, our method allows for real-time monitoring and optimization in cheese production, a feature not addressed in previous studies. The potential for cross-industry applications, such as fermentation monitoring in brewing and crystallization process optimization in pharmaceuticals, further underscores the versatility and impact of our approach.

7. Conclusion

Identifying the optimal curd time in cheese making is critical for enhancing production efficiency and ensuring consistent product quality. This process maximizes cheese yield and quality while minimizing whey fat losses.

In this study, we propose a novel, innovative, automated artificial intelligence-based method that leverages CV and ML techniques to automate the detection of curd-firming time during cheese production. Our method integrates sequence information from images with DEEP features, resulting in improved classification performance and leading to optimized cutting times, which are crucial for enhancing cheese quality and yield.

The release of CM-IDB, the first publicly available image dataset related to cheese-making, provides a valuable resource for future research and development in this field.

The experiments conducted in this study shed light on the performance of various classification methodologies when applied to single sets and merged sets. In particular, HC features exhibit moderate classification performance with notable variability across sets, while DEEP features consistently outperform HC features on average. Integrating sequence information with single image features, particularly the DEEP ones, through a spatiotemporal-based approach significantly enhances classification performance.

In light of the findings, the proposed approach appears well-suited for integration into real-time systems, notably owing to its swift inference times. This indicates its potential suitability for implementation in sectors like dairy manufacturing, where prompt and automated detection of curd-firming time holds significant importance.

Nevertheless, there are several avenues for enhancement. These include further exploring methods to amalgamate diverse features and implementing feature selection techniques to refine model performance. Additionally, investigating the impact of factors like photo backgrounds and illumination could significantly bolster the system's robustness.

Moreover, addressing the growing demand for transparency and accountability in artificial intelligence applications can lead to further studies that elucidate algorithmic decisions and predictions to ensure reliability and enhance end-user interpretability.

Future research includes expanding the research scope through cross-industry applications. For instance, the application of CV and ML techniques can be extended to monitor fermentation stages in beer brewing, optimize brewing times, and enhance product consistency. In pharmaceuticals, implementing these techniques during the crystallization process can improve the precision of drug production, ensuring consistent quality and efficacy. Additionally, using CV for agricultural monitoring can enhance crop growth tracking, early disease detection, and optimal harvest timing, significantly boosting productivity and minimizing losses.

Technological enhancements could also be pursued, such as integrating this approach with Internet of Things devices to facilitate real-time data collection and remote monitoring capabilities, thereby improving process control across various industries. Furthermore, investigating more sophisticated paradigms, such as reinforcement learning and generative adversarial network (GAN) architectures, could enhance the accuracy and reliability of curd-firming time detection and similar applications.

Finally, data diversification is another critical area for future research. Expanding the dataset to include a variety of cheese types and different production conditions can improve the model's robustness and generalizability. Additionally, employing techniques like GANs or diffusion models to generate synthetic data for rare or challenging scenarios can augment the dataset and enhance model training.

CRediT authorship contribution statement

Andrea Loddo: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Methodology, Investigation, Data curation, Conceptualization. **Cecilia Di Ruberto:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Investigation, Formal analysis, Data curation, Conceptualization. **Giuliano Armano:** Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Andrea Manconi:** Writing – review & editing, Validation, Supervision, Investigation, Formal analysis.

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.

Data availability

I have shared the link to the data in the manuscript and in the cover letter.

Acknowledgments

We would like to express our sincere gratitude to Gianluca Dettori of BiosAbbey s.r.l., to Massimiliano Sicilia of LAORE (Sardinian agency

Table A.13
Random Forest hyperparameters.

Parameter	Value
Number of Trees	100
Maximum Depth	10
Minimum Samples Split	2
Minimum Samples Leaf	1
Bootstrap	True
Criterion	Gini

Table A.14
Hyperparameters setup for the first MLP.

Parameter	Value
Hidden Layer Sizes	(20, 10)
Activation Function	ReLU
Solver	Adam
Alpha	0.0001
Learning Rate	Constant
Learning Rate Initialization	0.001
Maximum Iterations	200
Batch Size	Auto
Early Stopping	True
Momentum	0.9

Table A.15
Hyperparameters setup for the first MLP.

Parameter	Value
Hidden Layer Sizes	(80, 40, 20, 10)
Activation Function	ReLU
Solver	Adam
Alpha	0.0001
Learning Rate	Constant
Learning Rate Initialization	0.001
Maximum Iterations	200
Batch Size	Auto
Early Stopping	True
Momentum	0.9

for the implementation of regional agricultural and rural development programs), and to “Podda Formaggi” dairy for providing access to the images used in this research. Their contribution and expertise have been invaluable in simplifying and understanding the problem faced and enabling the analysis conducted in this study.

Appendix

In this section, the hyperparameters used for the experimental evaluation are comprehensively reported. In particular, Table A.13 shows the hyperparameters used to train the RF classifier, while Tables A.14 and A.15 present the hyperparameters chosen to train the two versions of MLP.

References

Alarcon, C., Shene, C., 2021. Fermentation 4.0, a case study on computer vision, soft sensor, connectivity, and control applied to the fermentation of a thraustochytrid. Comput. Ind. 128, 103431.
Arango, O., Castillo, M., 2018. A method for the inline measurement of milk gel firmness using an optical sensor. J. Dairy Sci. 101 (5), 3910–3917.
Armano, Giuliano, 2020. Using phidelta diagrams to discover relevant patterns in multilayer perceptrons. Sci. Rep. 10 (1), 21334.
Bellavista, Paolo, Bicocchi, Nicola, Fogli, Mattia, Giannelli, Carlo, Mamei, Marco, Picone, Marco, 2023. Requirements and design patterns for adaptive, autonomous, and context-aware digital twins in industry 4.0 digital factories. Comput. Ind. 149, 103918.
Bhatt, Rajesh, Naik, Naren, Subramanian, Venkatesh K., 2021. SSIM compliant modeling framework with denoising and deblurring applications. IEEE Trans. Image Process. 30, 2611–2626.
Bosakova-Ardenska, Atanaska, 2024. Recent trends in computer vision for cheese quality evaluation. Eng. Proc. 60 (1), 12.
Budelli, Eliana, Pérez, Nicolás, Negreira, Carlos, Lema, Patricia, 2017. Evaluation of ultrasonic techniques for on line coagulation monitoring in cheesemaking. J. Food Eng. 209, 83–88.
Budžaki, Sandra, Koceva Komleni’c, Daliborka, Lukinac Čači’c, Jasmina, Čači’c, F., Juki’c, M., Kožul, Ž., 2014. Influence of cookies composition on temperature profiles and qualitative parameters during baking. Croatian J. Food Sci. Technol. 6 (2), 72–78.
Deng, Jia, Dong, Wei, Socher, Richard, Li, Li-Jia, Li, Kai, Fei-Fei, Li, 2009. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. Ieee, pp. 248–255.
Di Ruberto, Cecilia, Putzu, Lorenzo, Rodriguez, Giuseppe, 2018. Fast and accurate computation of orthogonal moments for texture analysis. Pattern Recognit. 83, 498–510.
Dosovitskiy, Alexey, Beyer, Lucas, Kolesnikov, Alexander, Weissenborn, Dirk, Zhai, Xiaohua, Unterthiner, Thomas, Dehghani, Mostafa, Minderer, Matthias, Heigold, Georg, Gelly, Sylvain, Uszkoreit, Jakob, Houlsby, Neil, 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net.
Feng, Ran, Lillevang, Søren K., Ahrné, Lilia, 2021. Effect of water temperature and time during heating on mass loss and rheology of cheese curds. Foods 10 (11), 2881.
Galli, Bruno D., Hamed, Ahmed M., Sheehan, Jeremiah J., King, Niall, Abdel-Hamid, Mahmoud, Romeih, Ehab, 2023. Technological solutions and adaptive processing tools to mitigate the impact of seasonal variations in milk composition on Cheddar cheese production—A review. Int. J. Dairy Technol. 76 (3), 449–467.
Gao, Peng, Zhang, Wenyuan, Wei, Miaohong, Chen, Baorong, Zhu, Huiquan, Xie, Ning, Pang, Xiaoyang, Marie-Laure, Fauconnier, Zhang, Shuwen, Lv, Jiaping, 2022. Analysis of the non-volatile components and volatile compounds of hydrolysates derived from unmatured cheese curd hydrolysis by different enzymes. LWT 168, 113896.
Glorot, Xavier, Bengio, Yoshua, 2010. Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. JMLR Workshop and Conference Proceedings, pp. 249–256.
Guinee, Timothy P., 2021. Effect of high-temperature treatment of milk and whey protein denaturation on the properties of rennet–curd cheese: A review. Int. Dairy J. 121, 105095.
Haleem, Noman, Bustreo, Matteo, Del Bue, Alessio, 2021. A computer vision based online quality control system for textile yarns. Comput. Ind. 133, 103550.
Han, Kai, Wang, Yunhe, Chen, Hanting, Chen, Xinghao, Guo, Jianyuan, Liu, Zhenhua, Tang, Yehui, Xiao, An, Xu, Chunjing, Xu, Yixing, et al., 2022. A survey on vision transformer. IEEE Trans. Pattern Anal. Mach. Intell. 45 (1), 87–110.
Haralick, Robert M., Shanmugam, K., Dinstein, Its’Hak, 1973. Textural features for image classification. IEEE Trans. Syst. Man Cybern. SMC-3 (6), 610–621, Conference Name: IEEE Transactions on Systems, Man, and Cybernetics.
Hass, Roland, Munzke, Dorit, Vargas Ruiz, Salomé, Tippmann, Johannes, Reich, Oliver, 2015. Optical monitoring of chemical processes in turbid biogenic liquid dispersions by Photon Density Wave spectroscopy. Anal. Bioanal. Chem. 407, 2791–2802.
Haykin, Simon, Network, Neural, 2004. A comprehensive foundation. Neural Netw. 2 (2004), 41.
He, Dong-chen, Wang, Li, 1990. Texture unit, texture spectrum, and texture analysis. IEEE Trans. Geosci. Remote Sens. 28 (4), 509–512.
He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, Sun, Jian, 2016. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, IEEE, Las Vegas, NV, USA, pp. 770–778.
Hwang, Jeong Hyeon, Jung, Ah Hyun, Yu, Seung Su, Park, Sung Hee, 2022. Rapid freshness evaluation of cow milk at different storage temperatures using in situ electrical conductivity measurement. Innov. Food Sci. Emerg. Technol. 81, 103113.
Ionescu, Radu Tudor, Ionescu, Andreea Lavinia, Mothe, Josiane, Popescu, Dan, 2018. Patch autocorrelation features: a translation and rotation invariant approach for image classification. Artif. Intell. Rev. 49, 549–580.
Kamm, Simon, Veekati, Sushma Sri, Müller, Timo, Jazdi, Nasser, Weyrich, Michael, 2023. A survey on machine learning based analysis of heterogeneous data in industrial automation. Comput. Ind. 149, 103930.
Kang, Wenping, Liu, Shulin, Chen, Xiang, Feng, Kun, Guo, Zichen, Wang, Tao, 2022. Evaluation of ecosystem stability against climate changes via satellite data in the eastern sandy area of northern China. J. Environ. Manag. 308, 114596.
Lazouskaya, Maryna, Stulova, Irina, Sõrmus, Aavo, Scheler, Ott, Tiisma, Kalle, Vinter, Toomas, Loov, Roman, Tamm, Martti, 2021. Front-face fluorimeter for the determination of cutting time of cheese curd. Foods 10 (3), 576.
LeCun, Yann, Bengio, Yoshua, Hinton, Geoffrey, 2015. Deep learning. nature 521 (7553), 436–444.
Loddo, Andrea, 2024. CheeseMaking-IDB.
Loddo, Andrea, Di Ruberto, Cecilia, Armano, Giuliano, Manconi, Andrea, 2022. Automatic monitoring cheese ripeness using computer vision and artificial intelligence. IEEE Access 10, 122612–122626.
Moghiseh, Nasser, Arianfar, Akram, Salehi, Esmaeil Ataye, Rafe, Ali, 2021. Effect of inulin/kefiran mixture on the rheological and structural properties of mozzarella cheese. Int. J. Biiol. Macromol. 191, 1079–1086.

Mondini, Alessandro C., 2017. Measures of spatial autocorrelation changes in multitemporal SAR images for event landslides detection. Remote Sens. 9 (6), 554.

Moya-Sánchez, E. Ulises, Xambó-Descamps, Sebastià, Pérez, Abraham Sánchez, Salazar-Colores, Sebastián, Cortés, Ulises, 2021. A trainable monogenic ConvNet layer robust in front of large contrast changes in image classification. IEEE Access 9, 163735–163746.

Mukundan, R., Ong, S.H., Lee, P.A., 2001. Image analysis by Tchebichef moments. IEEE Trans. Image Process. 10 (9), 1357–1364, Conference Name: IEEE Transactions on Image Processing.

Ojala, Timo, Pietikainen, Matti, Maenpaa, Topi, 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intell. 24 (7), 971–987.

Panuju, Dyah R., Paull, David J., Griffin, Amy L., 2020. Change detection techniques based on multispectral images for investigating land cover dynamics. Remote Sens. 12 (11), 1781.

Peres, Ricardo Silva, Rocha, Andre Dionisio, Leitao, Paulo, Barata, Jose, 2018. IDARTS–towards intelligent data analysis and real-time supervision for industry 4.0. Comput. Ind. 101, 138–146.

Pieniazek, Facundo, Messina, Valeria, 2020. Microstructure, senescence and texture parameters of sardo cheese applying scanning electron microscopy with image analysis techniques. Microsc. Microanal. 26 (S1), 103–104.

Priyashantha, Hasitha, Höjer, Annika, Saedén, Karin Hallin, Lundh, Åse, Johansson, Monika, Bernes, Gun, Geladi, Paul, Hetta, Mårten, 2020. Use of near-infrared hyperspectral (NIR-HS) imaging to visualize and model the maturity of long-ripening hard cheeses. J. Food Eng. 264, 109687.

Putzu, Lorenzo, Di Ruberto, Cecilia, 2017. Rotation invariant co-occurrence matrix features. In: Battiato, Sebastiano, Gallo, Giovanni, Schettini, Raimondo, Stanco, Filippo (Eds.), Image Analysis and Processing - ICIAP 2017. In: Lecture Notes in Computer Science, Springer International Publishing, Cham, pp. 391–401.

Redmon, Joseph, Divvala, Santosh, Girshick, Ross, Farhadi, Ali, 2016. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 779–788.

Sabzi, Sajad, Arribas, Juan Ignacio, 2018. A visible-range computer-vision system for automated, non-intrusive assessment of the pH value in thomson oranges. Comput. Ind. 99, 69–82.

Søgaard, Jacob, Krasula, Lukáš, Shahid, Muhammad, Temel, Dogancan, Brunnström, Kjell, Razaak, Manzoor, 2016. Applicability of existing objective metrics of perceptual quality for adaptive video streaming. Electron. Imaging 28, 1–7.

Stocco, Giorgia, Summer, Andrea, Cipolat-Gotet, Claudio, Malacarne, Massimo, Cecchinato, Alessio, Amalfitano, Nicolò, Bittante, Giovanni, 2021. The mineral profile affects the coagulation pattern and cheese-making efficiency of bovine milk. J. Dairy Sci. 104 (8), 8439–8453.

Szegedy, Christian, Liu, Wei, Jia, Yangqing, Sermanet, Pierre, Reed, Scott, Anguelov, Dragomir, Erhan, Dumitru, Vanhoucke, Vincent, Rabinovich, Andrew, 2015. Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, IEEE Computer Society, pp. 1–9.

Szegedy, Christian, Vanhoucke, Vincent, Ioffe, Sergey, Shlens, Jon, Wojna, Zbigniew, 2016. Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, IEEE, Las Vegas, NV, USA, pp. 2818–2826.

Tabayehnejad, N., Castillo, M., Payne, F.A., 2012. Comparison of total milk-clotting activity measurement precision using the Berridge clotting time method and a proposed optical method. J. Food Eng. 108 (4), 549–556.

Taneja, Akriti, Nair, Gayathri, Joshi, Manisha, Sharma, Somesh, Sharma, Surabhi, Jambrak, Anet Rezek, Roselló-Soto, Elena, Barba, Francisco J., Castagnini, Juan M., Leksawasdi, Noppol, et al., 2023. Artificial intelligence: Implications for the agri-food sector. Agronomy 13 (5), 1397.

Tchuente, Dieudonné, Lonlac, Jerry, Kamsu-Foguem, Bernard, 2024. A methodological and theoretical framework for implementing explainable artificial intelligence (XAI) in business applications. Comput. Ind. 155, 104044.

Teague, Michael Reed, 1980. Image analysis via the general theory of moments*. J. Opt. Soc. Am. 70 (8), 920–930.

Teh, C.-H., Chin, R.T., 1988. On image analysis by the methods of moments. IEEE Trans. Pattern Anal. Mach. Intell. 10 (4), 496–513.

Tufano, Alessandro, Accorsi, Riccardo, Garbellini, Federica, Manzini, Riccardo, 2018. Plant design and control in food service industry. a multi-disciplinary decision-support system. Comput. Ind. 103, 72–85.

Vacca, Giuseppe M., Stocco, Giorgia, Dettori, Maria L., Bittante, Giovanni, Pazzola, Michele, 2020. Goat cheese yield and recovery of fat, protein, and total solids in curd are affected by milk coagulation properties. J. Dairy Sci. 103 (2), 1352–1365.

Wang, Zhou, Bovik, Alan C., 2009. Mean squared error: Love it or leave it? A new look at signal fidelity measures. IEEE Signal Process. Mag. 26 (1), 98–117.

Wang, Zhou, Bovik, Alan C., Sheikh, Hamid R., Simoncelli, Eero P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. 13 (4), 600–612.

Xu, Yaqing, Qamsane, Yassine, Puchala, Saumuy, Januszczak, Annette, Tilbury, Dawn M., Barton, Kira, 2024. A data-driven approach toward a machine-and system-level performance monitoring digital twin for production lines. Comput. Ind. 157, 104086.

Yuan, Junfang, Bian, Zhengfu, Yan, Qingwu, Pan, Yuanqing, 2019. Spatio-temporal distributions of the land use efficiency coupling coordination degree in mining cities of western China. Sustainability 11 (19), 5288.