K. Trejo et al. (Eds.)

doi:10.3233/FAIA250613

Decoding Cheese Ripening: A Deep Learning Approach to Non-Invasive Monitoring

Daniel PARDO [a,b,1], Mehmet Oguz MULAYIM [a,b] and Manuel CASTILLO [b]
[a] *Artificial Intelligence Research Institute, IIIA-CSIC*
[b] *Universitat Autònoma de Barcelona, UAB*

Abstract. In cheese production, the ripening phase is crucial to developing the final product's sensory characteristics. However, non-invasive and real-time monitoring remains a challenge. To address this, computer vision techniques offer a promising opportunity for more precise, automated, and continuous tracking of the ripening process. This work proposes a deep-learning-based methodology to classify cheese wheels by ripening day, providing valuable insights for quality control and process standardization, which are critically needed in the cheese industry. Using digital images, we assess the feasibility of a Machine Learning system to estimate ripening stages visually. As a proof of concept, we employ pre-trained Convolutional Neural Networks to extract visual descriptors for classification. Unlike previous studies relying solely on embeddings from pre-trained models, this work fine-tunes these models to adapt the visual representation to the specific domain of cheese ripening, allowing a comparative analysis of both approaches.

Keywords. Cheese production, Cheese ripening, Computer vision, Deep learning, Non-invasive monitoring

1. Introduction

The ripening phase is the final stage in the cheese production process and occurs in most cheeses, with the exception of fresh cheeses. This stage can last from a few weeks to even years and is when sensory characteristics such as the flavor, texture, or aroma of each cheese variety develop. Different microorganisms and enzymes induce a complex series of biochemical and microbiological events that must be controlled to achieve the desired characteristics [1].

Determining the degree of ripening of a cheese is crucial to understanding the state of development of its sensory characteristics and is associated with the quality of the cheese. These sensory attributes can be evaluated using different techniques and methods, classified into three large groups: *non-destructive instrumental techniques* such as infrared spectroscopy or optical techniques that have potential for online monitoring; *conventional sensory analysis* based on consumer or expert panels to evaluate the characteristics once the process has finished; *chemical analysis* of metabolites and volatile

[1] Corresponding Author: Daniel Pardo, daniel.pardo@iiia.csic.es

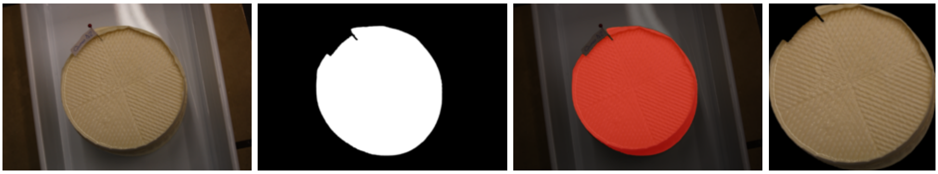
compounds using techniques such as Gas Chromatography and Mass Spectrometry (GC-MS), Gas Chromatography and Olfactometry (GC-O) or Solid Phase MicroExtraction (SPME) [2].

In this context, the application of Machine Learning (ML) techniques to the cheese ripening process has gained attention in recent years, particularly when combined with spectrometry, computer vision, and volatile compound analysis. Most studies have aimed to predict sensory attributes or estimate the ripening level, often expressed as the number of days or weeks since the start of maturation. The development of non-invasive methods opens new opportunities for real-time online control, with computer vision standing out as a particularly promising approach. For example, [3,4] analyzed digital images of cheese wheels at various ripening stages to classify either the exact day within a time window or whether the cheese had reached the desired maturity level or not. In both studies, pre-trained Convolutional Neural Network (CNN)s were used to extract visual features and train classification models.

However, these pre-trained models on ImageNet [5] do not contain the cheese class and have not been specifically tuned to the cheese domain, suggesting that domain-specific adaptation could further improve performance. In this work we want to explore whether adjusting the models to the cheese domain can help improve the models. To this end, we fine-tuned existing CNNs models using a dataset of cheese images across different ripening days and evaluate the impact on classification accuracy.

2. Dataset and preprocessing

We used the [6] dataset, which contains 217 images of cheese wheels at a resolution of 3872 × 2592 pixels. The images correspond to six cheeses obtained from two different milk batches and were labeled during a 60-day ripening process. To improve the model's learning capacity, we performed a semi-automatic image segmentation to isolate the wheel region and remove the background as shown in Figure 1. We applied the Segment Anything Model (SAM) [7], giving one point at the center of each image as the segmentation prompt. The resulting masks were refined with morphological operations such as openings and closings, and their circularity was automatically checked to keep the best one. Finally, the segmented region was cropped and stored as a new image in 224 × 224 pixels. At training time, each RGB image was channel-wise normalized with the ImageNet mean and standard deviation.



(a) Original image (b) Final mask (c) Image segmented (d) Final image

Figure 1. Image preprocessing steps.

3. Methodology

To simplify the classification task, the samples were grouped into three classes according to ripening days: [0–20], [21–40] and [41–60], with roughly equal counts per class (76, 72, and 69, respectively). The images come from six cheeses produced in two production batches (milk batches A and B). Therefore, samples inside the same batch cannot be considered independent.

To avoid data leakage, the training and test sets were split by batch of origin: batch A (109 samples) was used only for training and validation, while batch B (108 samples) was used for testing. In batch B, one cheese is larger than the rest; for this reason, this batch was chosen for testing to evaluate the model's generalization capability in a more robust way.

We used the models ResNet-101 [8] and EfficientNetV2-M [9] pretrained on ImageNet and available in torchvision [2]. From each model we extracted the embeddings from the global average-pooling layer (2048 and 1280 features, respectively). These embeddings are numerical vectors that represent the features and patterns learned by the model from the input data. We then used them as input to train three classifiers: Multi Layer Perceptron (MLP), Support Vector Machines (SVM), and eXtreme Gradient Boosting (XGBoost).

We evaluated three strategies to train both CNN models, comparing them by their accuracy in predicting the ripening period: (i) Direct use of the original weights; (ii) Adding a new head layer with the backbone (the feature extractor part of the model) frozen, updating BatchNorm running statistics; and (iii) Fine-tuning of the last convolutional block, keeping the rest of the backbone frozen.

To select the best hyper-parameter configuration in strategies ii and iii, we applied 5-fold stratified cross-validation. Once the best configuration was identified, the model was retrained with all the training data and saved to obtain the final embeddings. Likewise, each of the three classifiers was optimized with 5-fold stratified cross-validation on the training embeddings and finally fitted using the optimal hyperparameters on the whole set. The test set was used only once, at the end of the process, to compare the performance of the different models.

4. Results and Discussion

The results show that domain adaptation is a key factor in improving performance. In our specific case of cheese wheels, the networks pre-trained on ImageNet do not inherently learn the cheese category; thus, explicitly feeding this information to the models enhances their generalization capabilities. The results are presented in Table 1. In all cases, the embeddings obtained after domain-adaptation (strategies ii and iii) outperform those generated with the original weights, highlighting the importance of domain adaptation. Overall, the ResNet-101 backbone achieves better metrics than EfficientNet V2-M, except when the latter is combined with XGBoost. Strategy iii consistently increases test accuracy for all three classifiers, although it exhibits overfitting (training accuracy close to 100%). In contrast, strategy ii offers slightly lower absolute performance but a more balanced train-test compromise.

[2] https://docs.pytorch.org/vision/stable/index.html

Table 1. 5-fold cross-validation accuracy (mean ± standard deviation) and test accuracy for the three fine-tuning strategies. Best test score in **bold**.

Backbone	**Classifier**	**Strategy i**		**Strategy ii**		**Strategy iii**	
		Train	Test	Train	Test	Train	Test
ResNet-101	MLP	0.59 ± 0.15	0.52	0.67 ± 0.17	0.62	0.97 ± 0.04	0.60
EfficientNet V2-M	MLP	0.53 ± 0.14	0.43	0.61 ± 0.15	0.47	0.96 ± 0.03	0.58
ResNet-101	SVM	0.85 ± 0.08	0.51	0.83 ± 0.08	0.61	1.00 ± 0.00	**0.70**
EfficientNet V2-M	SVM	0.79 ± 0.08	0.44	0.74 ± 0.07	0.53	0.97 ± 0.04	0.65
ResNet-101	XGBoost	0.78 ± 0.09	0.44	0.74 ± 0.10	0.49	0.98 ± 0.02	0.63
EfficientNet V2-M	XGBoost	0.80 ± 0.06	0.49	0.76 ± 0.08	0.52	0.88 ± 0.08	0.60

As for classification algorithms, the SVM achieved the best performance, obtaining the highest average validation accuracy and the best test accuracy (70%) when using strategy iii. The standard deviation between folds is low ($\sigma \leq 0.08$), suggesting a stable estimate of system performance despite the small training set size. The optimum performance of SVM was achieved with hyperparameters $C = 1$ and $\gamma = 0.1$ using an RBF kernel, indicating that a non-linear distance-based decision frontier fits the embedding distribution better than MLP layers or XGBoost trees.

Table 2 shows the confusion matrices corresponding to the best-performing configuration (ResNet-101 as backbone and SVM as classifier) under the three different strategies. We observe that with strategies i and ii, the model tends to overpredict class 2, the middle temporal segment. In contrast, strategy iii produces a more balanced distribution across all classes, suggesting better generalization performance.

Table 2. Confusion matrices for the three strategies

(a) Strategy i

		Predicted		
		C1	C2	C3
Actual	C1	17	19	3
	C2	6	29	4
	C3	2	19	9

(b) Strategy ii

		Predicted		
		C1	C2	C3
Actual	C1	20	17	2
	C2	1	34	4
	C3	1	17	12

(c) Strategy iii

		Predicted		
		C1	C2	C3
Actual	C1	25	12	2
	C2	0	29	10
	C3	2	6	22

5. Conclusion and Future work

This work confirms that applying fine-tuning to a pre-trained model increases generalization capabilities for estimating cheese ripening days: accuracy on the test batch increased from 52% (original pre-trained weights, strategy i) to 70% after fine-tuning the last block of ResNet-101. This significant improvement highlights the importance of domain adaptation in this problem, as ImageNet lacks cheese examples. Therefore, due to the limitations of the original dataset, pre-trained models can learn general features such as texture, color, or lighting from images, but they cannot understand the variability inherent to cheese without this additional tuning.

Further research is warranted to overcome the limitations in this study. Among the most significant are the size and low variability of the dataset, the reduction of the problem to 20-day intervals, and the exploration of a limited number of architectures and classifiers. In addition to addressing these limitations, we propose analyzing alternative methodological approaches such as deploying an end-to-end network, applying data augmentation, or using more complex transformer-based models. All this is aimed at improving the results obtained and addressing cheese ripening prediction with a finer temporal resolution, ideally at the day level.

Acknowledgments

Daniel Pardo is funded by the AGAUR FI Joan Oró grant (2024 FI-1 00424). Mehmet Oguz Mulayim is funded by European Union KHealthInAir and GUARDEN projects with No. 101057693 and 101060693. This work was supported by grant PID2022-136787NB-I00 funded by MCIN/AEI/10.13039/501100011033. Manuel Castillo is funded by pHotoYOGURT (ACC_2023_EXP_SIA001_40_0000092; DACC, GENCAT) and OPTICmizing (2023 INNOV 00027; AGAUR, GENCAT) projects.

References

[1] Fox PF, Guinee TP, Cogan TM, McSweeney PLH. Biochemistry of Cheese Ripening. In: Fox PF, Guinee TP, Cogan TM, McSweeney PLH, editors. Fundamentals of Cheese Science. Boston, MA: Springer US; 2017. p. 391-442. Available from: https://doi.org/10.1007/978-1-4899-7681-9_12.

[2] Khattab AR, Guirguis HA, Tawfik SM, Farag MA. Cheese ripening: A review on modern technologies towards flavor enhancement, process acceleration and improved quality assessment. Trends in Food Science & Technology. 2019 Jun;88:343-60. Available from: https://www.sciencedirect.com/science/article/pii/S0924224419300597.

[3] Loddo A, Di Ruberto C, Armano G, Manconi A. Automatic Monitoring Cheese Ripeness Using Computer Vision and Artificial Intelligence. IEEE Access. 2022;10:122612-26. Conference Name: IEEE Access. Available from: https://ieeexplore.ieee.org/document/9956763/?arnumber=9956763.

[4] Zedda L, Perniciano A, Loddo A, Di Ruberto C. Understanding cheese ripeness: An artificial intelligence-based approach for hierarchical classification. Knowledge-Based Systems. 2024 Jul;295:111833. Available from: https://www.sciencedirect.com/science/article/pii/S0950705124004672.

[5] Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. Ieee; 2009. p. 248-55.

[6] Perez-Playà B, Verdugo-González L, Pardo D. Cheese Ripening RGB Image Dataset; 2025. Available from: https://zenodo.org/records/15298940.

[7] Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, et al.. Segment Anything. arXiv; 2023. ArXiv:2304.02643 [cs]. Available from: http://arxiv.org/abs/2304.02643.

[8] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. arXiv; 2015. ArXiv:1512.03385 [cs]. Available from: http://arxiv.org/abs/1512.03385.

[9] Tan M, Le QV. EfficientNetV2: Smaller Models and Faster Training. arXiv; 2021. ArXiv:2104.00298 [cs]. Available from: http://arxiv.org/abs/2104.00298.