eISSN 1981-1829

Ciência e Agrotecnologia
Science and Agrotechnology

Comparing human and machine clustering for tomato ripening stage classification

Comparando agrupamento humano e de máquina para classificação do estádio de maturação do tomate

Erick Rodríguez Hernández[1], Juan Carlos Olguin Rojas[2], Gerardo Antonio Alvarez Hernandez[3*], Juan Irving Vasquez-Gomez[3], Abril Valeria Uriarte Arcia[3], Hind Taud[3]

ABSTRACT

The classification of tomato ripening stages involves assigning a tomato to a category based on the visual indicators of its maturity. Indeed, the specific number of categories and their attributes are determined by the agricultural standards of each country, which rely on an empirical understanding of visual characteristics. Conversely, automatic unsupervised classification techniques, such as deep learning-based methods, autonomously learn their characteristics. In this research, a comparison is made between expert-based classification and unsupervised classification, with a particular focus on the analysis of the number of clusters and their respective features. Remarkably, this investigation finds an alignment in the number of clusters identified by both methods. This discovery supports the notion that the expert-based classification system is compatible with automated approaches. The outcomes of this research could aid the agricultural sector in refining automatic classification techniques. Furthermore, this work provides the scientific community with valuable insights into the clustering of images by machine learning methods

Index terms: Precision farming; deep learning; unsupervised learning; clustering.

RESUMO

A classificação do estágio de maturação do tomate envolve atribuir uma classe a um tomate com base nos aspectos visuais da sua maturidade. De fato, o número de classes e suas características são definidos pelas normas do departamento agrícola de cada país, as quais são baseadas no conhecimento empírico sobre as características visuais. Por outro lado, métodos de classificação automática não supervisionados, como aqueles baseados em aprendizado profundo, aprendem suas características de forma independente. Neste trabalho, comparamos a classificação baseada em especialistas com a classificação não supervisionada, analisando particularmente o número de agrupamentos (clusters) e as características de cada um. Surpreendentemente, nosso estudo revela uma coincidência no número de agrupamentos para ambas as abordagens. Com essa descoberta, fornecemos evidências de que a classificação atual baseada em especialistas também é adequada para métodos automáticos. As descobertas deste estudo podem ajudar a indústria agrícola a desenvolver métodos precisos de classificação automática. Além disso, para a comunidade científica, este estudo oferece percepções sobre como os métodos de aprendizado de máquina agrupam imagens.

Termos para indexação: Agricultura de precisão; aprendizagem profunda; aprendizagem não supervisionados; agrupamento.

Introduction

Recent advancements in computer vision have enabled the execution of numerous precision agriculture tasks outdoors with considerable accuracy. Such tasks include fruit detection (Sharif et al., 2018; Yuan et al., 2020; Ale et al., 2019; Liu et al., 2020; Liu , Pi & Xia, 2020), fruit segmentation (Shiu, Lee, & Chang, 2023; Zhu et al., 2023; Fujinaga & Nakanishi, 2023; Jia et al., 2022), three-dimensional reconstruction (Tao & Zhou, 2017; Jun et al., 2021; Chen, et al., 2020; Yandun, Silwal, & Kantor, 2020; Louedec, Li, & Grzegorz, 2020), and grasp planning (Rong et al., 2022; Guo et al., 2020). The foundation of many of these techniques is the artificial intelligence paradigm known as connectionist, particularly through supervised learning with deep neural networks (DNNs). Despite their efficiency in addressing tasks, these methods are not without limitations. They necessitate substantial data volumes and frequently do not provide clear justifications for their decisions. These limitations serve as the impetus for exploring how algorithmic decisions align with

Agricultural Sciences

Ciênc. Agrotec., 48:e019123, 2024
http://dx.doi.org/10.1590/1413-7054202448019123

Editor: Renato Paiva

[1]Continental Autonomous Mobility, Santiago de Querétaro, Querétaro, México
[2]Universidad Autónoma Chapingo/UACh, Departamento de Ingeniería Mecánica Agrícola/DIMA, Texcoco, Estado de México, México
[3]Centro de Innovación y Desarrollo Tecnológico en Cómputo/CIDETEC, Instituto Politécnico Nacional/ IPN, Ciudad de Mexico, Ciudad de Mexico, México

*Corresponding author: galvarezh1400@alumno.ipn.mx

Received in December 5, 2023 and approved in March 25, 2024

human reasoning, with a specific focus on a case study: the classification of tomato ripening stages.

The tomato represents one of the most widely consumed vegetables in everyday human diets, with its consumption reaching millions of individuals daily. In Mexico, the tomato stands out as a critical fruit due to its demand, value, and production volume, ranking it among the top ten tomato producers globally with a production of 3,461,766 tons. However, escalating labor costs are increasingly becoming a constraint in numerous agricultural sectors. This paper concentrates on the classification of ripening stages, which is predominantly conducted manually in most of the industry.

The stages of ripening delineate the maturity level of a tomato over time, providing essential information for farmers to determine the best time for harvest and to make other well-informed decisions. While the criteria for ripening stages may differ from one country or region to another in Mexico (where this research is situated), they are based on empirical guidelines established in the Mexican Norm. This approach is referred to as human expertise, wherein classification relies on observable characteristics of the tomato. These guidelines will be elaborated upon later in the document. In contrast, DNN-based methods utilize their unique feature extraction mechanisms, shaped by the architecture of the neural network and the optimization strategies employed. The characteristics that are considered important by humans may vastly differ from those identified by neural networks. Furthermore, the number of stages recognized by humans might substantially diverge from the number of stages discerned by a computer. Thus, this study aims to investigate whether human expertise in classification aligns with state-of-the-art deep learning-based classification.

During our review of the literature, various studies were found to propose supervised methods for fruit classification (Chen, Cheng, & Liu, 2022; Chen et al., 2022). However, supervised classification essentially seeks to mimic the knowledge of experts, particularly through the labels they assign. This limitation has led us to employ unsupervised classification, a method that autonomously identifies groups to minimize intra-group differences while maximizing inter-group differences. By adopting unsupervised classification, we aim to compare algorithmic decision-making with human expertise, focusing not only on the features of tomatoes but also on the optimal number of groups as determined by the algorithm.

For the comparative analysis, a dataset consisting of 3,000 images was assembled. These images underwent a conversion to a latent representation, mirroring the process utilized by state-of-the-art neural networks. Subsequently, unsupervised classification was applied to determine the optimal number of clusters based on the sum of squared errors. The results of this clustering were then compared with the stages delineated by human experts. Remarkably, the analysis disclosed a correspondence in the number of stages identified by both methods, with a similar set of features being recognized.

The outcomes underscore the compatibility of existing expert-based classification frameworks with the development of accurate automated techniques within the agricultural sector. Furthermore, this research offers the scientific community profound insights into the image clustering processes employed by machine learning algorithms.

Material and Methods

Related Work

Fruit classification is an important task in agriculture. In the last decade, it has been automated by the use of appropriate methods and technologies. For example, nondestructive tools, such as colorimeters, visible and near-infrared (VNIR) spectroscopy (Walsh et al., 2020), hyperspectral imaging (Su et al., 2021), visible imaging (Zhang et al., 2018), fluorescent imaging (Matveyeva et al., 2022) and electronic noses (Baietto & Wilson, 2015), have been employed for the collection of characteristics that allow fruit sorting, mainly for light sensors in the visible and nonvisible range.

Image-based fruit classification systems have been developed to reduce reliance on multiple sensors and lower implementation costs in real agricultural fields, utilizing artificial intelligence models. These models employ various machine learning approaches, including supervised learning, where an expert in fruit classification provides the correct categorization of fruits through labels. This enables the model to learn the necessary features for accurate classification. For instance, Chen et al. (2021) introduced a fruit classifier using a multiple optimization convolutional neural network (MC-CNN), optimizing the weights from training on a virtual convolutional network that classifies using a Self-Organizing Map (SOM) network. This approach significantly enhances fruit generalization, achieving a classification accuracy of 99%.

Other methods for fruit classification use convolutional neural networks (CNN) and optimization algorithms. Chen, Cheng, and Liu (2022) developed a system for fruit quality classification that combines fruit detection using YOLO-Tiny with a condition classification using their custom CNN, reaching 88% accuracy in fruit detection. Another approach by Chen et al. (2022) for detecting citrus ripeness involves a two-step process: initial detection with the YOLOv5 model and ripening state classification using a 4-channel ResNet34 network enhanced by a visual saliency algorithm, achieving an average accuracy of 95.07%. Alharbi et al. (2023) introduced a fruit classifier using an AFC-ETSAFDL technique, which employs a fusion-based feature extraction method with three Deep Learning (DL) models—DenseNet, ResNet, and

Inceptionv3—optimized with an ETSA. Classification is performed using the Extreme Gradient Boosting (XGBoost) model, obtaining an average accuracy of 95.68% using 70% of the TR database. Conversely, Hernandez et al. (2023) applied the YOLOv3-tiny object detector for classifying tomato ripening stages into six categories, optimizing the model through a search grid to determine the best hyperparameters. This detector achieved an average F1-score of 90.0%. Appe, Arulselvi, and Balaji (2023) modified YOLOv5 for tomato ripening state classification by adding a convolutional block attention module (CAM) to the model to improve its accuracy. In addition, to prevent object repetition and overlapping, they added nonmaximal suppression and distance of intersection over junction (DIoU), which they named CAM-YOLO. This model reported an average accuracy of 88.1% and was capable of detecting small and overlapping tomatoes.

Another type of learning is unsupervised learning, which consists of giving the model only the images of the fruit to be classified and allowing the model to infer the corresponding classes and perform the classification with respect to these. This type of learning is less common, but several examples, such as Zhang and Xu (2018), proposed using an unsupervised image segmentation algorithm based on conditional random fields (ULCRF) to perform segmentation to obtain the best-located object to be subsequently classified. They reported that they achieved better classification results using this automatic segmentation method than using supervised learning methods. Knott, Perez, and Defraeye (2023) proposed a Vision-Transformer to perform fruit classification but with little data. They reported that it is able to achieve 90% classification accuracy faster than a CNN-based classifier.

In addition to the aforementioned learning types, there exists the potential to enhance results through the integration of supervised and unsupervised learning methods. For instance, Xue, Liu, and Ma (2023) introduced a hybrid approach for fruit classification that merges both supervised and unsupervised learning. This innovative method employs a generative adversarial network (GAN) to create synthetic images of fruits. These images are then utilized to train a convolutional neural network (CNN) for the task of classification.

In summary, numerous studies employ supervised learning techniques for fruit classification, wherein these studies replicate human knowledge. Conversely, a limited number of studies focus on unsupervised learning; yet, they fail to offer details regarding the distribution of classes. Furthermore, these unsupervised methods are not suitable for classifying tomato ripening stages.

Method overview

In this research, we compared human expert-based clustering of tomato ripening data to unsupervised clustering. To achieve our objectives, we follow the methodology depicted in Figure 1, which is explained next.

Initially, a dataset comprising 3,000 RGB images was collected. This dataset includes images of tomatoes at various maturity stages. Each image within the dataset shows a single tomato against a black background to minimize experimental variability. Two methodologies were then applied to this dataset: one using human expertise and the other utilizing contemporary machine clustering techniques. In the first methodology, images were manually classified into a predetermined number of clusters, referred to as ripening stages, based on the empirical knowledge outlined in the Mexican standard. Detailed information on the

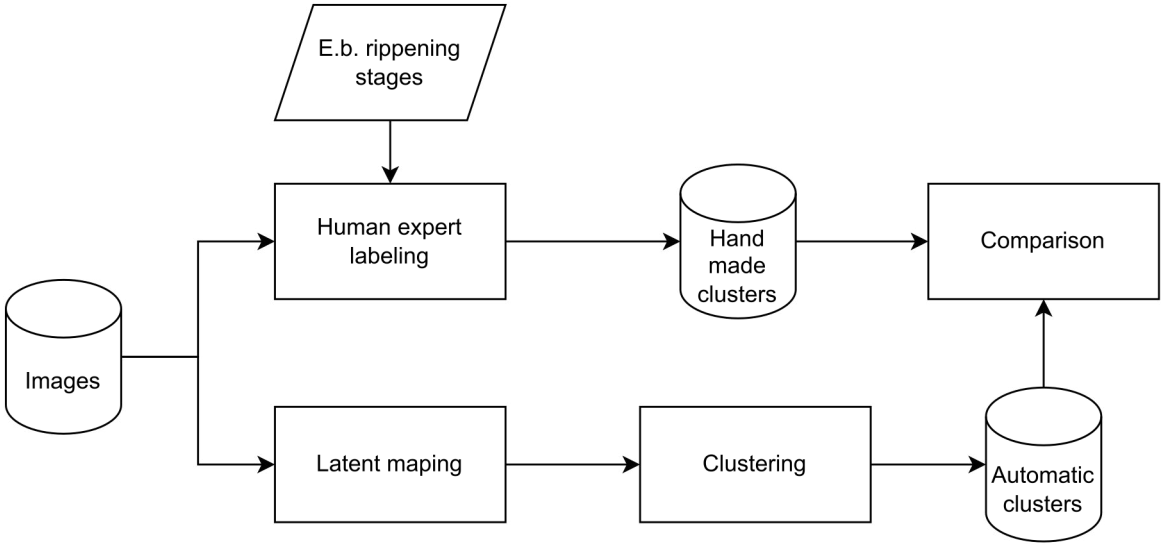


Figure 1: Methodology for comparing the human expert-based (E.B.) clustering of tomato ripening versus a data driven clustering.

expert-based classification of ripening stages is presented in the section "Tomato Dataset Collection." Following this, an organized dataset was created by categorizing the images into clusters. In the alternate methodology, the collected images were transformed into their latent representations. This transformation involved converting the image matrices into a more compact representation using a pre-trained encoder. The rationale behind this conversion is to retain only the primary abstract features of the images in this reduced space. The latent representations were then clustered through unsupervised learning. Afterward, a new dataset was formed based on the automatically generated clusters. In conclusion, the clusters formed by each methodology were compared to address the research question concerning the extent of similarity between human expert clustering and the algorithm-driven approach.

Tomato dataset collection

The tomatoes were harvested from a greenhouse situated in Chignahuapan, Puebla, Mexico. A greenhouse was chosen for this case study due to its prevalent use in tomato cultivation. Following the harvest, the tomatoes were manually sorted based on their coloration, and each sorted batch was subsequently photographed. To capture the RGB images, a specialized photography setup was constructed with controlled lighting conditions. Figure 2 illustrates the components of this setup, which comprises a Logitech C920 camera capable of capturing images at a maximum resolution of 1080p and 30 fps. A mobile platform powered by a 5 V DC gear motor with a no-load speed of 200 rpm and a torque of 800 g/cm was mounted as a base to move the tomato plant. At 10 cm from the base, an aluminum dome was installed to diffuse the light produced by the LED ring, achieving evenly distributed lighting. This was placed opposite the base, ensuring that the light did not directly reach the fruit or be captured by the camera. For the background of the prototype, a black backdrop was used to minimize noise caused by light in the scene captured by the camera. In total, 500 images were captured for each of the six categories considered by the official Mexican Standard NMX-FF-031-1997, resulting in a total of 3,000 images of tomatoes with dimensions of 800 × 600 pixels.

Human expert-based clustering

Tomatoes can be harvested at different stages of maturity, either at physiological maturity or commercial maturity; the choice depends on the market's needs. The color of the tomato is the easiest indicator for defining the stage of maturity, transitioning from green coloration (indicative of immaturity) to red coloration (indicative of maturity), as shown in Figure 3.

In this research, the six maturation stages as defined by the Official Mexican Standard NMX-FF-031-1997 were considered. These maturation stages include green, breaker, turning, orange, orange-red, and red, as shown in Figure 4. Typically, tomatoes intended for export are harvested during the initial three stages, ranging from green to turning, to ensure a minimum shelf life of two weeks under appropriate temperature conditions. Similarly, a classification chart by the USDA (United States Department of Agriculture) distinguishes six maturation stages based on color, naming them green, breaker, turning, pink, light red, and red. However, the analysis of this regulation is left for future work.

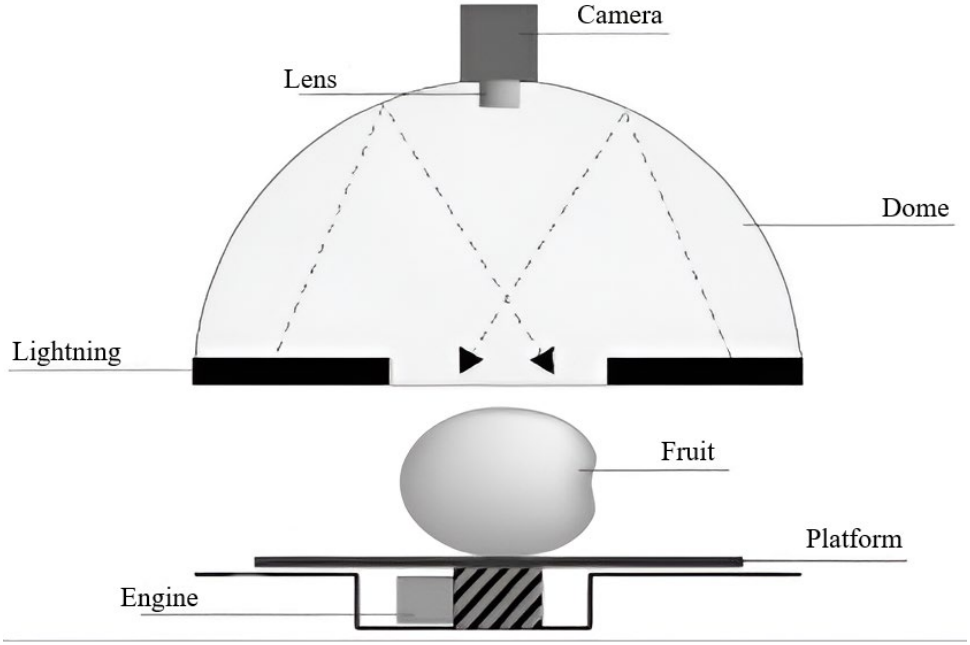


Figure 2: Image shooting mechanism. The mechanism includes a rotating platform with controlled illumination to capture different points of view of the tomato.

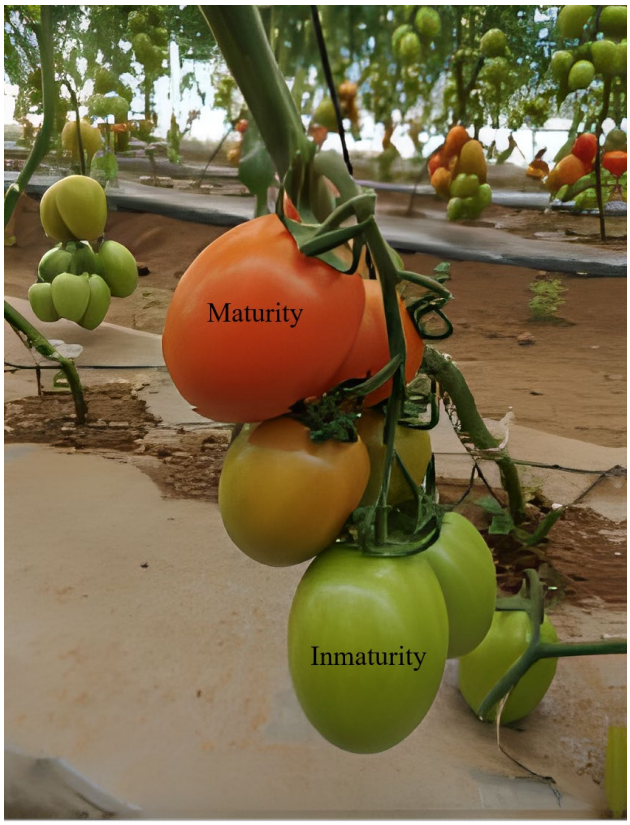


Figure 3: Tomato fruits with different degrees of coloration at maturity.

Once the tomatoes were collected and grouped into six categories, they were photographed using an image-shooting mechanism. Several pictures of each tomato were taken, resulting in 500 pictures for each category. Then, the pictures are stored in separate folders to be compared against the automatic clustering.

Automatic clustering

As illustrated in Figure 1, the process of automatic clustering begins by mapping the images to a latent space. In deep learning, a latent space represents a condensed form of data abstracted or encoded from high-dimensional input data, such as HD images, into a lower-dimensional format, such as shorter vectors. This principle is commonly applied in unsupervised learning models like autoencoders and generative adversarial networks (GANs), with the objective being to learn a representation of the input data that encapsulates its most critical or "latent" characteristics (Goodfellow, Bengio, & Courville, 2016). This mapping technique facilitates the simplification of the automatic clustering process by concentrating on the high-level features rather than addressing all photometric variables present in the images.

In this study, the mapping is performed using a convolutional feature extractor, which constitutes a section of a convolutional network that comprises solely the convolution operations, excluding the fully connected layers. The VGG16 (Simonyan & Zisserman, 2015) feature extractor is employed for this purpose. The convolution kernels were set through transfer learning, where the source task involved the classification of the ImageNet dataset.

Once the images are transformed into latent vectors, the K-means method is employed for clustering. The objective is to arrange the data in such a way that points within each cluster are closely grouped together while maintaining a considerable distance from points in other clusters. In K-means, this objective is achieved by minimizing the variance within each cluster, which is quantified as the sum of squared distances (SSD) between each point and the centroid of its respective cluster.

However, a limitation of the K-means method is that the number of clusters must be predetermined. To reduce human intervention in determining the number of clusters, K, a two-step analysis is conducted. Initially, the elbow method is applied. The basic idea behind the elbow method is to run K-means across a range of cluster numbers (k) and measure the performance for each value of k. In this implementarion, the performance is measured with the SSE (Sum of Squared Errors). Then the point of maximum curvature is identified by examining the plot and looking for a point where the rate of decrease in the performance metric sharply changes, resembling an "elbow." The idea is that

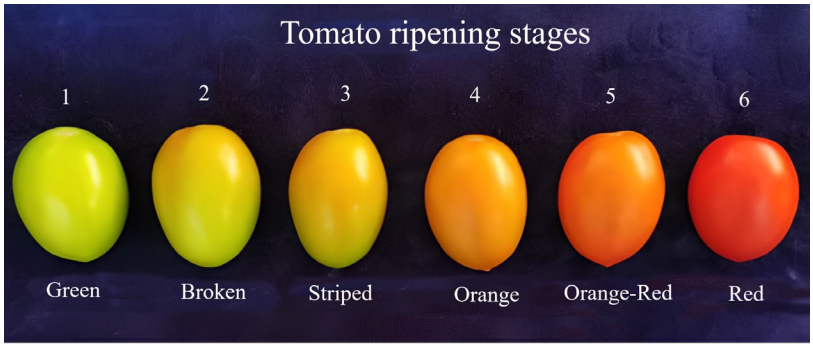


Figure 4: The six stages of maturation indicated in the official Mexican standard NMX-FF-031-1997.

adding more clusters beyond this point does not provide much better modeling of the data, as the decrease in the performance metric starts to level off. Usually only the elbow method output is taken as the optimal number of cluster, however, we go beyond and we use additional information. In the second step, the silhouette coefficient is calculated for each candidate. The silhouette coefficient is a metric used to evaluate the quality of clusters created by the k-means clustering. It measures how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette coefficient provides insight into the distance between the resulting clusters and the density of the clusters themselves. To combine both methods, the local maximum of the silhouette coefficient nearest to the value indicated by the elbow method is then selected. This approach enables the automatic computation of the optimal number of clusters for the K-means method.

Results and Discussion

In this section, we will present the results of the experiments and analyze them based on three key aspects: the number of clusters automatically formed, the distribution of points in the latent space, and the features characterizing each automatic cluster. Furthermore, we will conduct a comparative analysis between these automatic clusters and those identified by human experts.

Number of clusters

Initially, we examined the determination of the number of clusters. In this study, we commenced with a predetermined range of clusters spanning from 1 to 40. Subsequently, both the SSE and the Silhouette coefficient were computed for each value within this range. The summarized outcomes are depicted in Figure 5. From this preliminary investigation, it was observed that beyond 22 clusters, the Silhouette coefficient fluctuates, and its value diminishes compared to a previous number of clusters, such as 11 centroids. Consequently, we opted to confine the experiment to a narrower range, spanning from 0 to 15 clusters. Within this revised range, the experiment was repeated.

The outcomes of testing the number of clusters from 1 to 15 are illustrated in Figure 6. These results indicate a decrease in the SSE, as anticipated, when employing the elbow method. Conversely, the silhouette coefficient demonstrates less fluctuation. Consequently, from these findings, a particular number of clusters was determined. To ascertain this number, we

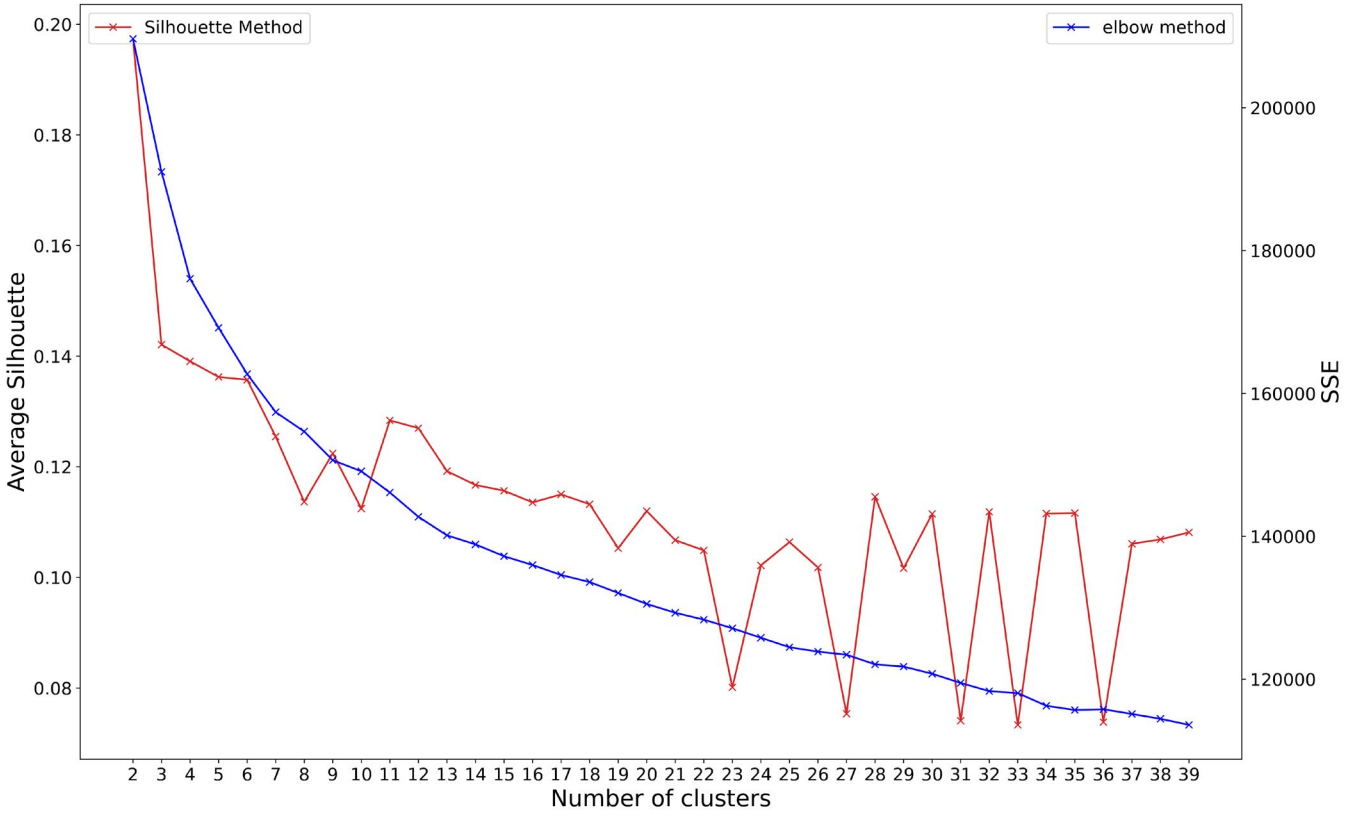


Figure 5: Graph of the SSE and coefficient silhouette vs. number of clusters for clusters ranging from 0 to 40. We observe that the average silhouette oscillates from the SSE curve after 22 clusters.

applied the elbow method to the 15 calculations. The resulting number of clusters was identified as six, a determination supported by the silhouette coefficient, where six corresponds to a local maximum.

Analysis

Our findings have uncovered a remarkable convergence in the number of clusters (six) for determining the ripening stages. This alignment between the human expert-based stages and automatic clustering is unexpected, considering that the automatic method extracts its own features, which are not defined by human knowledge. We initially anticipated a different number of clusters based on these features. Consequently, these results underscore the efficacy of the current human rule, as six stages are not only discernible by humans but also by computers.

To offer a deeper insight into the criteria adopted by the automatic clustering, we have undertaken the task of "reconstructing" the centroids of each cluster using a decoder trained to replicate the input of the feature extractor. The methodology for centroid reconstruction involves training an autoencoder based on the U-Net architecture (Ronneberger, Fischer, & Brox, 2015). An autoencoder comprises two components: an encoder and a decoder. In this scenario, the encoder corresponds to the feature extractor utilized in this study, namely, the VGG16 feature extractor as detailed in the section "Automatic clustering." Subsequently, the decoder is constructed to mirror the encoder but in reverse. The autoencoder is then trained using all the images in the dataset, where the input is an image and the output is compared to the input using mean squared error (MSE). Once the autoencoder training is completed, only the decoder is utilized by inputting the centroids of the formed clusters to obtain reconstructed images. The resulting images are displayed in Figure 7.

As observed in Figure 7, the color variation between centroids is a significant factor in class identification. The centroid colors range from green (Cluster 5) to red (Cluster 4). Shape represents a second factor; despite tomatoes belonging to the same species, there is a noticeable variation in shape among the centroids across clusters. Specifically, Cluster 1 exhibited a more rounded shape than the other clusters. Consequently, the automatic clustering algorithm anticipates minor shape variations from the initial to the final maturity stages.

To provide a deeper analysis of the changes in color between clusters, we plotted the RGB histograms for each cluster in Figure 8. We observe from the histograms that the clusters with more green are clusters five, six, and two, while the clusters with more red are clusters four and three.

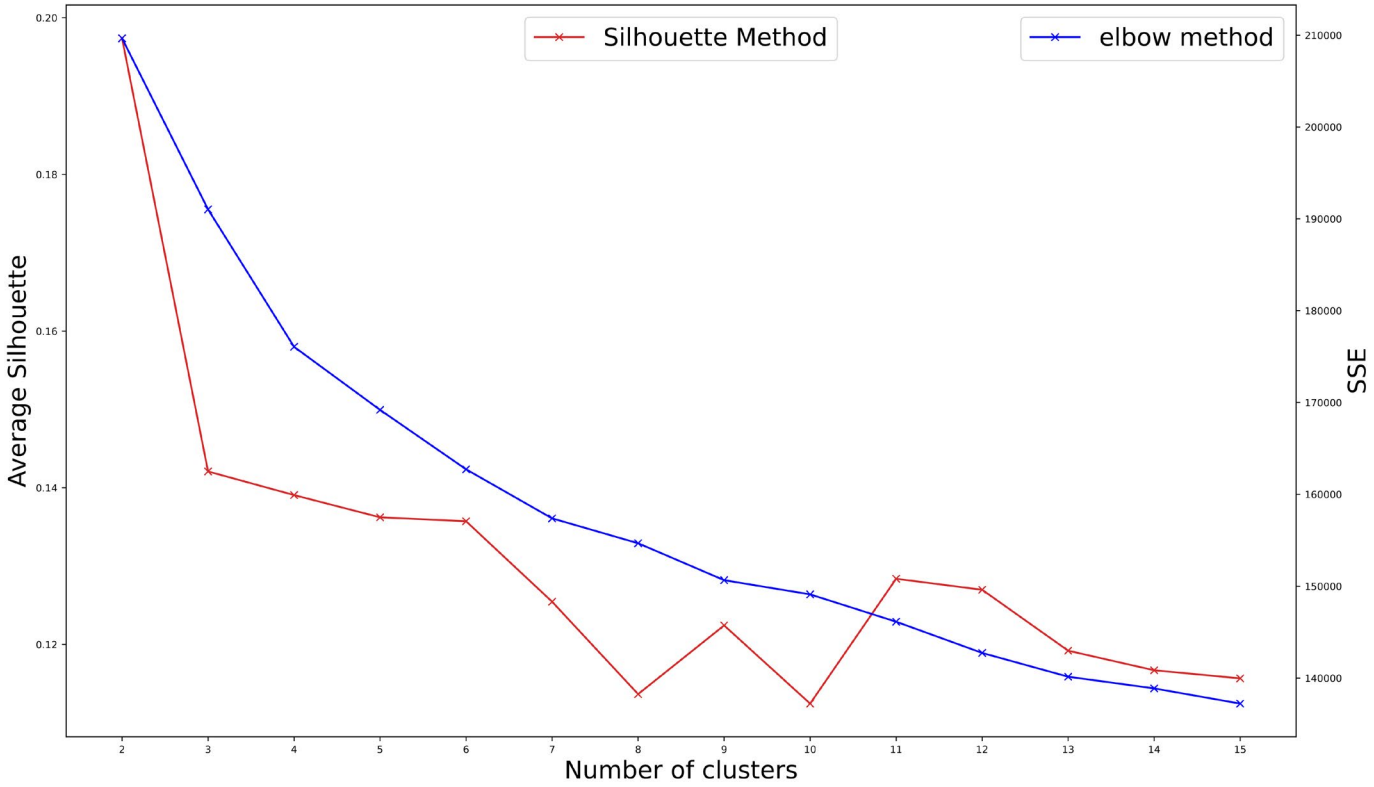


Figure 6: In the graph of the SSE vs. number of clusters, the inflection point of the curve is at 6 according to the elbow method and in combination with the silhouette method.

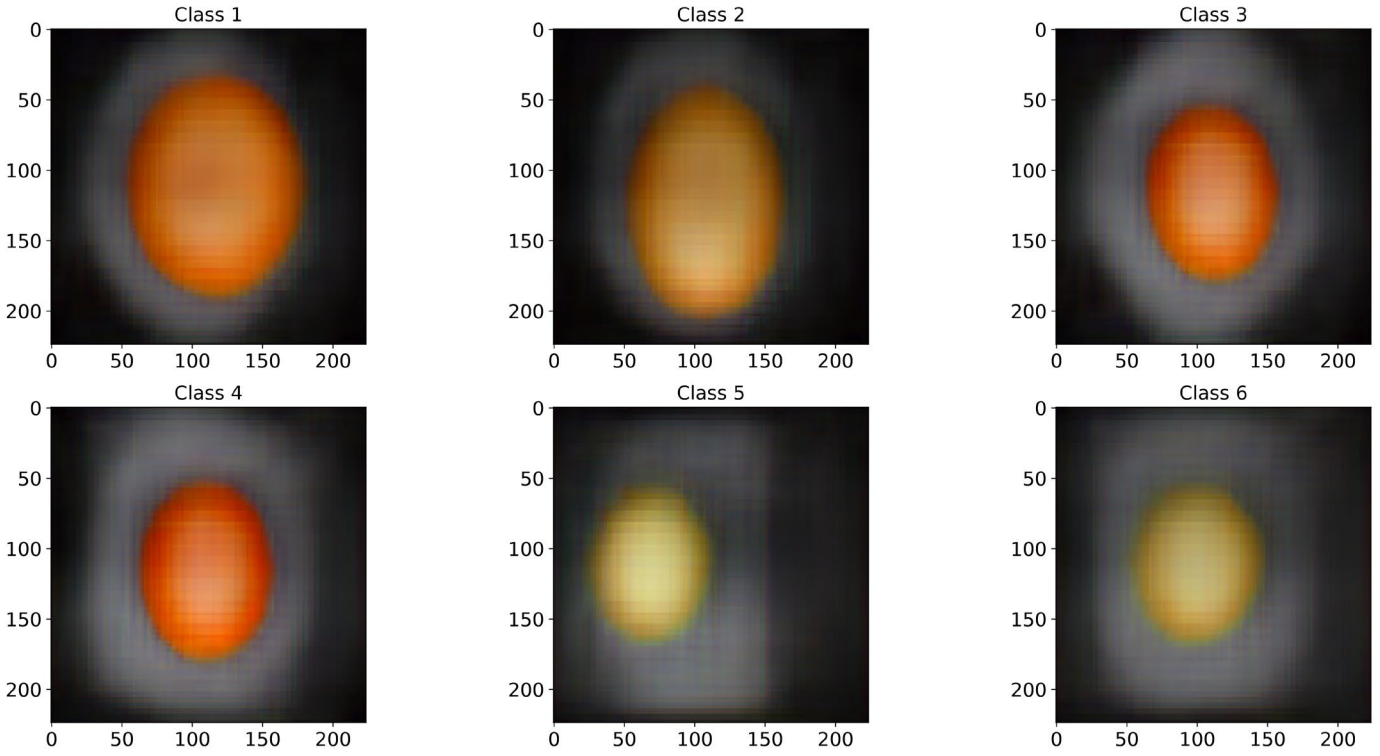


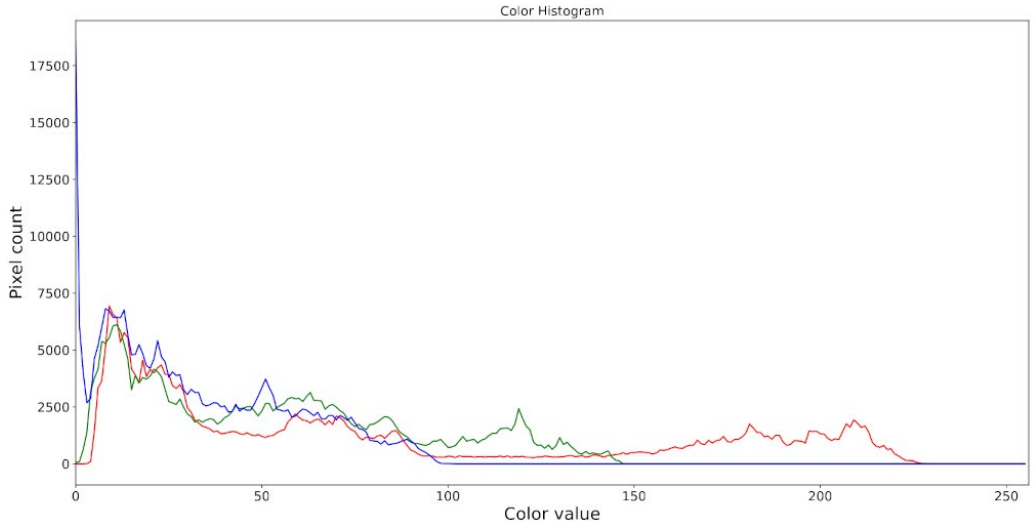
Figure 7: Illustration of the reconstructed images originating from the centroids of each class according to the k-means algorithm. These images are synthetic and represent the 'mean' image for each formed cluster. Thus, they contain the mean features of each cluster.

We can utilize this information to compare automatic clustering with human-expert clustering. Firstly, it is notable that in both cases, the number of clusters is six. This indicates that it is equally manageable for both humans and machines to classify the degree of maturity into six stages, thereby addressing the first question of our experimental design. Moreover, we observe that the color of the tomato serves as a crucial feature, as evidenced by the reconstruction of the centroids where color serves as a distinguishing factor between classes. This aspect fulfills the third question of our experimental design, highlighting the significance of color in both scenarios. However, for the automatic method, shape is also taken into account.

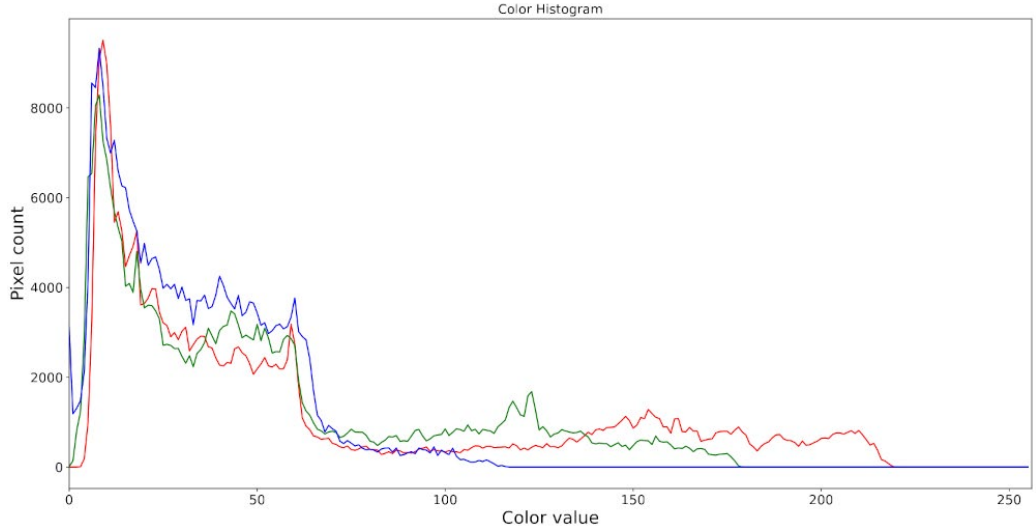
To address the final question regarding the distribution of latent points, we applied the t-SNE method to map the latent space points to a 2-dimensional plot. This step was essential for visually representing the points, as multidimensional points cannot be directly visualized. The resulting plot is depicted in Figure 9. Despite the limitations of the mapping technique, we can observe that the clusters are visually grouped. Notably, the fifth cluster appears notably distant from the others, presumably representing the green class. Conversely, the first cluster appears to be dispersed among the others, a phenomenon that can be elucidated by its centroid being situated between the red and green classes.

It is essential to note that this experiment was conducted under controlled conditions, and the conclusions drawn are valid within this controlled environment. To extend the applicability of the conclusions, it is imperative to incorporate images captured under diverse conditions. However, to the best of our knowledge, there are no other public databases with the same classes. Therefore, this aspect of the study is left for future work.

(a) Centroid 1



(b) Centroid 2



(c) Centroid 3

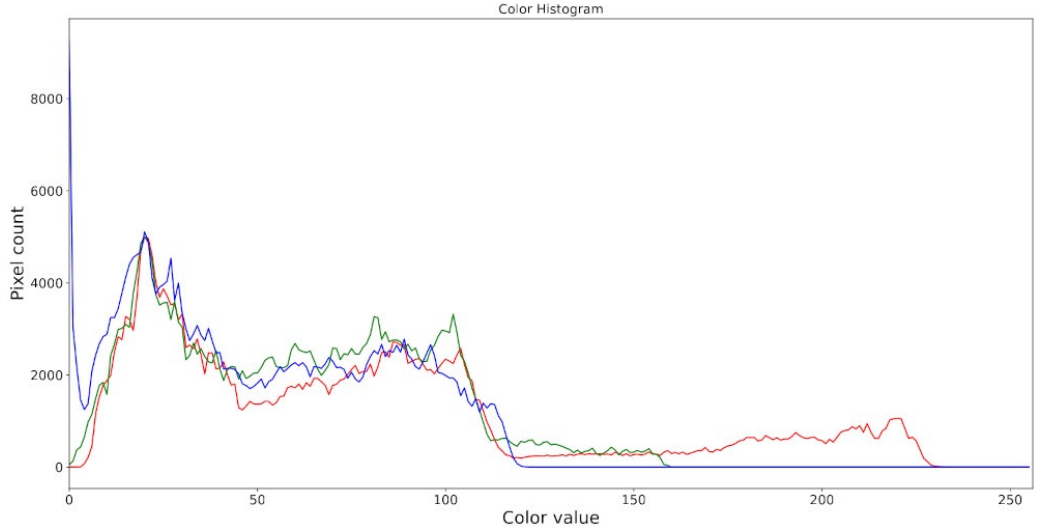
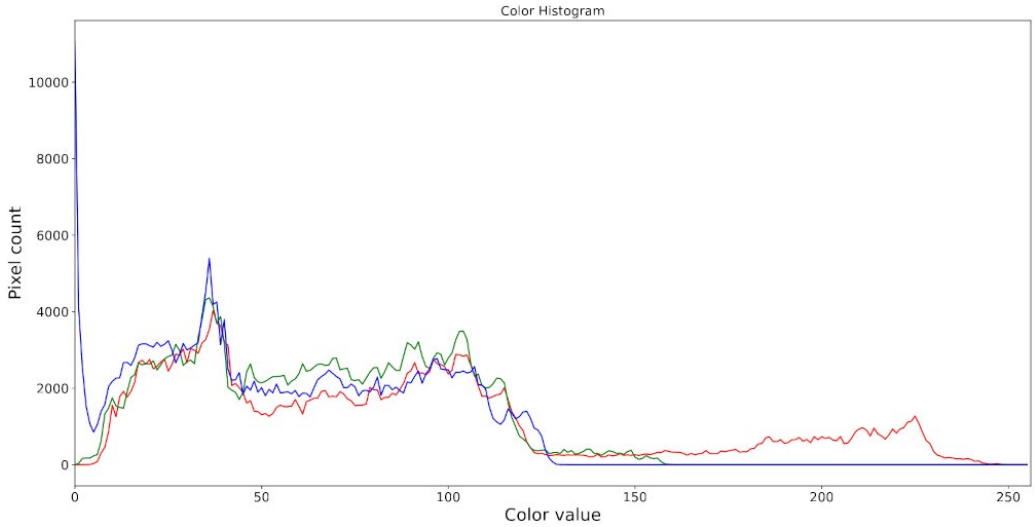
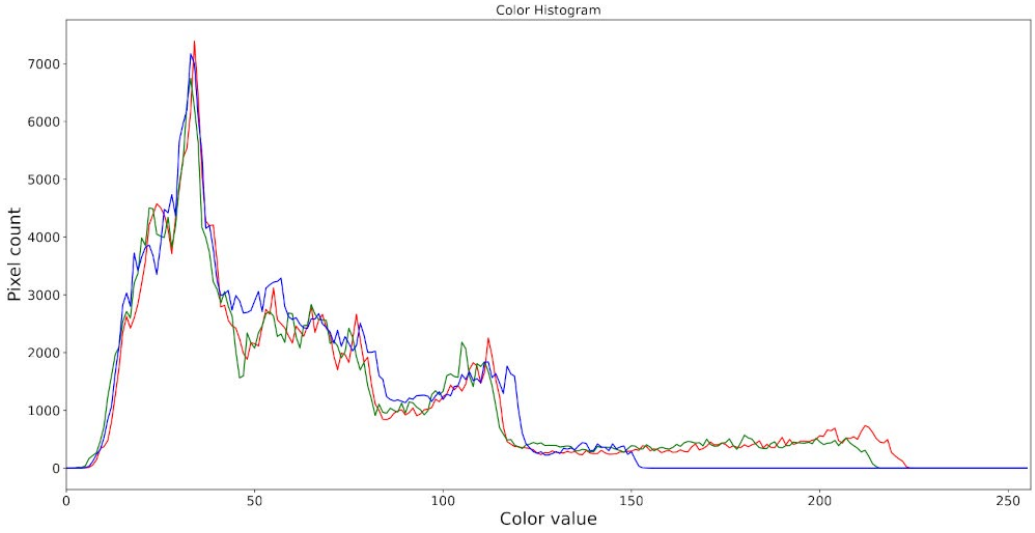


Figure 8: Continuation.

(d) Centroid 4



(e) Centroid 5



(f) Centroid 6

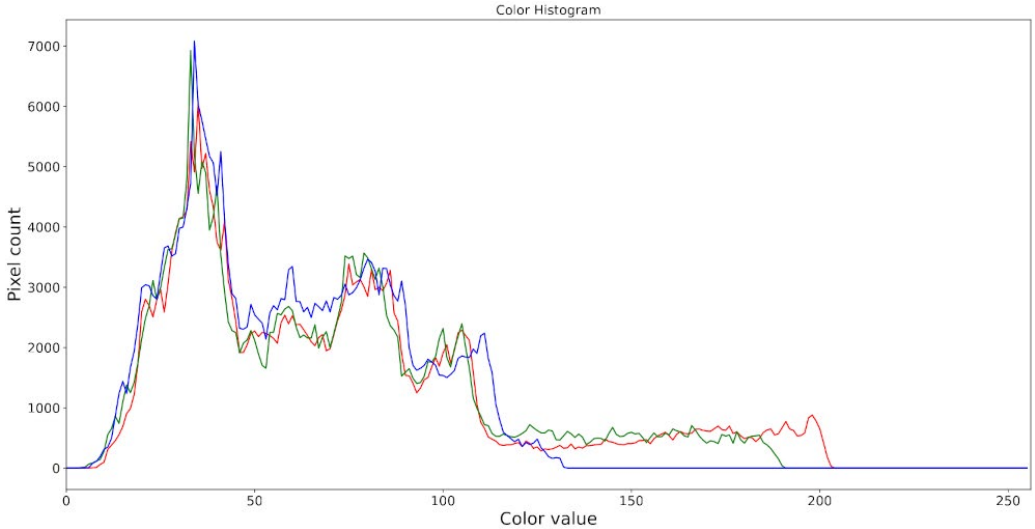


Figure 8: Histogram of the reconstructed centroids from each cluster. We can observe how the colors vary across the clusters. Each subfigure provides the histogram for each color channel. With these figures, we can analyze the color variation.

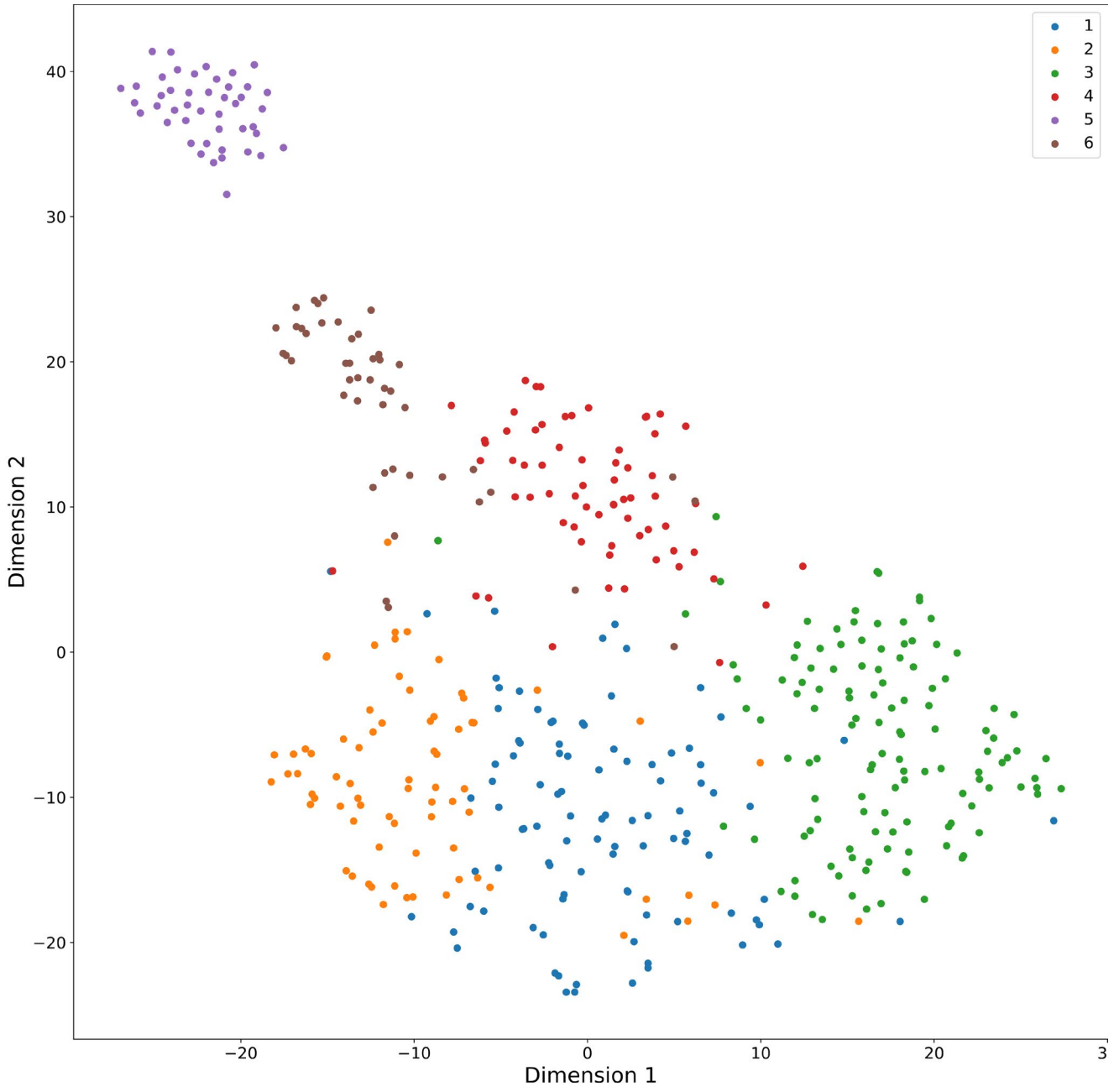


Figure 9: 2D representation of the embedding using t-SNE and color according to each class of tomatoes clustered in an unsupervised manner. The number assigned to each color corresponds to each class of tomatoes. We can observe that the clusters are relatively well separated. Only some nonlinear confusion is observed between clusters 1 and 2.

Conclusions

A comparative study between the human expert-based determination of ripening stages and automatic clustering has been presented. The automatic approach identified six clusters as the optimal number, a finding that remarkably aligns with the human expert-based classification. Hence, we conclude that this number is suitable for developing automatic applications. Furthermore, we provide evidence indicating that the primary feature learned by the automatic method is color, as it serves as a distinct discriminator among the centroids.

Acknowledgments

The authors thank the labor of Alfonso Martínez Guevara for taking the pictures. We thank the support of Project SIP IPN 20231897 and CONACYT SNI.

Author Contribution

Conceptual idea: Vasquez, J.I.; Rodríguez Hernandez E. Methodology design: Vasquez, J.I; Uriarte Arcia A. Data collection: Olguín Rojas J.C., Alvarez Hernández G., Data analysis and interpretation: Taud H.; Vasquez, J.I.; Rodríguez Hernandez E.; Writing and editing: Vasquez, J.I., Rodríguez Hernandez E., Uriarte Arcia A., Olguín Rojas J.C., Taud H.

References

Ale, L. et al. (2019). Deep learning based plant disease detection for smart agriculture. *IEEE Globecom Workshops*, 1-6.

Alharbi, A. H. et al. (2023). Automated fruit classification using enhanced tunicate swarm algorithm with fusion based deep learning. *Computers and Electrical Engineering*, *108*:108657.

Appe, S. N., Arulselvi, G., & Balaji, G. (2023). Cam-yolo: Tomato detection and classification based on improved yolov5 using combining attention mechanism. *PeerJ Computer Science*, *9*:e1463.

Baietto, M., Wilson, A. D. (2015). Electronic-nose applications for fruit identification, ripeness and quality grading. *Sensors*, *15*(1):899-931.

Chen, M. et al. (2020). Three-dimensional perception of orchard banana central stock enhanced by adaptive multi-vision technology. *Computers and Electronics in Agriculture*, *174*:105508.

Chen, S. et al. (2022). Citrus fruits maturity detection in natural environments based on convolutional neural networks and visual saliency map. *Precision Agriculture*, *23*:1515-1531.

Chen, M. C., Cheng, Y. T., & Liu, C. Y. (2022). Implementation of a fruit quality classification application using an artificial intelligence algorithm. *Sensors & Materials*, *34*(1):151-162.

Chen, X. et al. (2021). The fruit classification algorithm based on the multi-optimization convolutional neural network. *Multimedia Tools and Applications*, *80*:11313-11330.

Fujinaga, T., & Nakanishi, T. (2023). Semantic segmentation of strawberry plants using deeplabv3+ for small agricultural robot. *IEEE/SICE International Symposium on System Integration (SII)*, 1-6.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning, MIT press*. 800p.

Guo, N. et al. (2020). Pose estimation and adaptable grasp configuration with point cloud registration and geometry understanding for fruit grasp planning. *Computers and Electronics in Agriculture*, *179*:105818.

Hernandez, G. A. A. et al. (2023). Detection of tomato ripening stages using yolov3-tiny. *ArXiv preprint arXiv:2302.00164*, 1-12.

Jia, W. et al. (2022). Polar-net: Green fruit example segmentation in complex orchard environment. *Frontiers in Plant Science*, *13*:5176.

Jun, J. et al. (2021). Towards an efficient tomato harvesting robot: 3D perception, manipulation, and end-effector. *IEEE access*, *9*:17631-17640.

Knott, M., Perez, C. F., & Defraeye, T. (2023). Facilitated machine learning for image-based fruit quality assessment. *Journal of Food Engineering*, *345*:111401.

Le Louedec, J., Li, B., & Cielniak, G. (2020). *Evaluation of 3D vision systems for detection of small objects in agricultural environments*. In VISIGRAPP (5: VISAPP), pp. 682-689.

Liu, G. et al. (2020). Yolo-tomato: A robust algorithm for tomato detection based on YOLOv3. *Sensors*, *20*(7):2145.

Liu, J., Pi, J., & Xia, L. (2020). A novel and high precision tomato maturity recognition algorithm based on multi-level deep Residual network. *Multimedia Tools and Applications*, *79*(13):9403-9417.

Matveyeva, T. A. et al. (2022). Using fluorescence spectroscopy to detect rot in fruit and vegetable crops. *Applied Sciences*, *12*(7):3391.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention*, *18*:234-241.

Rong, J. et al. (2022). Fruit pose recognition and directional orderly grasping strategies for tomato harvesting robots. *Computers and Electronics in Agriculture*, *202*:107430.

Sharif, M. et al. (2018). Detection and classification of citrus diseases in agriculture based on optimized weighted segmentation and feature selection. *Computers and Electronics in Agriculture*, *150*:220-234.

Shiu, Y. S., Lee, R. Y., & Chang, Y. C. (2023). Pineapples' detection and segmentation based on faster and mask R-CNN in UAV imagery. *Remote Sensing*, *15*(3):814.

Simonyan, K., & Zisserman, A. (2015). *Very deep convolutional networks for large-scale image recognition*. 3rd International Conference on Learning Representations (ICLR 2015). 1-14.

Su, Z. et al. (2021). Application of hyperspectral imaging for maturity and soluble content determination of strawberry with deep leaning approaches. *Frontiers in Plant Science*, *12*:736334.

Tao, Y., & Zhou, J. (2017). Automatic apple recognition based on the fusion of color and 3D feature for robotic fruit picking. *Computers and Electronics in Agriculture*, *142*:388-396.

Walsh, K. B. et al. (2020). Visible-nir 'point'spectroscopy in postharvest fruit and vegetable assessment: The science behind three decades of commercial use. *Postharvest Biology and Technology*, *168*:111246.

Xue, G., Liu S.; Ma, Y. (2023). A hybrid deep learning-based fruit classification using attention model and convolution autoencoder. *Complex and Intelligent Systems*, *9*:2209-221.

Yandun, F., Silwal, A., & Kantor, G. (2020). Visual 3D reconstruction and dynamic simulation of fruit trees for robotic manipulation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 54-55.

Yuan, T. et al. (2020). Robust cherry tomatoes detection algorithm in greenhouse scene based on SSD. *Agriculture*, *10*(5):160.

Zhang, P. et al. (2018). Infrared and visible image fusion using co-occurrence filter. *Infrared Physics and Technology*, *93*:223-231.

Zhang, P., & Xu, L. (2018). Unsupervised segmentation of greenhouse plant images based on statistical method. *Scientific Reports*, *8*(1):4465.

Zhu, W. et al. (2023). Segmentation and recognition of filed sweet pepper based on improved self-attention convolutional neural networks. *Multimedia Systems*, *29*(1):223-234.