Article

A Multi-Considered Seed Coat Pattern Classification of *Allium* L. Using Unsupervised Machine Learning

Gantulga Ariunzaya [1], Shukherdorj Baasanmunkh [2], Hyeok Jae Choi [2], Jonathan C. L. Kavalan [3] and Sungwook Chung [1,*]

[1] Department of Computer Engineering, Changwon National University, Changwon 51140, Republic of Korea
[2] Department of Biology and Chemistry, Changwon National University, Changwon 51140, Republic of Korea
[3] Department of Computer and Information Science and Engineering, University of Florida, Gainesville, FL 32611, USA
* Correspondence: swchung@changwon.ac.kr; Tel.: +82-55-213-3819

Abstract: The seed coat sculpture is one of the most important taxonomic distinguishing features. The objective of this study is to classify coat patterns of *Allium* L. seeds into new groups using scanning electron microscopy unsupervised machine learning. Selected images of seed coat patterns from more than 100 *Allium* species described in literature and data from our samples were classified into seven types of anticlinal (irregular curved, irregular curved to nearly straight, straight, S, U, U to Ω, and Ω) and five types of periclinal walls (granule, small verrucae, large verrucae, marginal verrucae, and verrucate verrucae). We used five unsupervised machine learning approaches: K-means, K-means++, Minibatch K-means, Spectral, and Birch. The elbow and silhouette approaches were then used to determine the number of clusters required. Thereafter, we compared human- and machine-based results and proposed a new clustering. We then separated the data into six target clusters: SI, SS, SM, NS, PS, and PD. The proposed strongly identical grouping is distinct from the other groups in that the results are exactly the same, but PD is unrelated to the others. Thus, unsupervised machine learning has been shown to support the development of new groups in the *Allium* seed coat pattern.

Keywords: *Allium* seed coat; testa sculpture; unsupervised machine learning; SEM; new grouping

check for updates

Citation: Ariunzaya, G.; Baasanmunkh, S.; Choi, H.J.; Kavalan, J.C.L.; Chung, S. A Multi-Considered Seed Coat Pattern Classification of *Allium* L. Using Unsupervised Machine Learning. *Plants* **2022**, *11*, 3097. https://doi.org/10.3390/plants11223097

Academic Editor: Božena Šerá

Received: 17 October 2022
Accepted: 10 November 2022
Published: 14 November 2022

1. Introduction

The genus *Allium* L. is the largest genera in the family Amaryllidaceae [1–3] and consists of more than 1000 species [4] distributed in the northern hemisphere and southern Africa [1,4]. The most recent *Allium* taxonomy consists of 800 species, 15 subgenera, and 56 subsections [1]. Several research groups worldwide are studying *Allium* species. Depending on their macro- and micro-morphological, molecular information, biogeographical distribution, and evolutionary history, some ambiguous subgenus and sections have been updated [1]. Therefore, scientists and researchers have begun to investigate the morphological, anatomical, developmental, and phytochemical characteristics of a variety of *Allium* taxa.

Seed coat pattern is one of the most important taxonomic features that support taxonomic relationships in *Allium* [5–10]. The species level typically exhibits a high degree of diversity, and infraspecific diversity is also significant [8–11]. According to Yusupov et al. [3], the macro- and micro-morphology of seeds are one of the most important taxonomic features that delimit the taxa in *Allium*. Moreover, Celep et al. [9] identified testa cell form, shape, and sculpturing of periclinal walls, and the position, shape, and type of undulation of anticlinal walls as essential diagnostic features used to categorize taxa at the sectional level. Seed testa sculpture attributes in combination with seed shape provide key features for distinguishing major clades of *Allium* in molecular phylogeny [1]. In general, there are two types of walls: anticlinal wall undulation includes irregularly curved, irregularly curved to nearly straight, Ω, U, U to Ω, S, and straight, whereas periclinal wall undulation includes granule, small verrucae, large verrucae, central verrucae, marginal verrucae, and dense granule [5–16].

Recently, identification and classification based on deep learning and machine learning have greatly influenced several research disciplines. One of them is the field of plants. Some studies use plant disease and flower classifications already trained using convolutional neural networks (CNNs) and transfer learning to conduct new studies and improve the productivity of previous works. For example, Geetharamani [17] used a trained CNN to identify an open dataset of plant leaf diseases with 39 different classes. He used CNN to extract valuable leaf features directly from the raw representations of the input data and used a deconvolutional network technique to gain an understanding of the selected characteristics [18]. Reyes et al. [19] pre-trained a CNN using 1.8 million images and applied a fine-tuning technique to transfer the learning recognition capabilities from general domains to the unique issue of plant identification. Researchers prefer machine-learning-based techniques over human-based classification because deep learning and machine learning have several advantages with specific data.

In recent decades, machine learning and deep learning have been successfully applied in biology for classification, segmentation, and identification. For example, Saleem et al. [20] proposed two steps to classify plant diseases using a CNN. First, they compared well-known CNN architectures with modified and cascaded/hybrid versions of some of the deep-learning models proposed in recent works to determine the CNN that performed best. They then evaluated how well the best model performed after training with different deep-learning optimizers. In addition, Wang et al. [21] used real-time detection of tomato diseases in the environment based on YOLOv3-tiny network architecture. These research methodologies are distinguished by their use of labeled datasets. These datasets are used to train or monitor algorithms to accurately identify data or predict outcomes. With labeled input and outputs, the model can measure its accuracy and learn over time. In contrast, we looked at unsupervised machine learning, where unlabeled datasets are analyzed and clustered. The main idea is to discover hidden patterns in the data without requiring human intervention.

In recent years, researchers have become interested in plant identification and classification through machine and deep learning. Several studies used machine-learning and deep-learning classifiers to classify plant diseases by processing leaf and seed images. Most researchers used support vector machines to classify grape plant diseases [22,23], as well as neural networks [24]. In addition, the deep-learning model of the Efficient model can classify plant leaf disease [25,26]. Deep learning was used to classify tomato leaf diseases [27], and banana leaf diseases [28]. However, these studies are only based on the classification of plant diseases using machine learning such as deep learning.

The identification and classification of various seed coat patterns using machine learning and deep learning have not been fully studied. Similar to our study, a new class of coralline algae was created by fine-tuning pre-trained CNNs [29]. They classified four commonly occurring Mediterranean species (*Lithothamnion corallioides*, *Mesophyllum philippii*, *Lithophyllum racemus*, and *Lithophyllum pseudoracemus*) at the species and genus levels. Therefore, we describe how to determine our own grouping based on the seed coat pattern using machine learning.

In this study, we used unsupervised machine-learning techniques to classify a set of scanning electron microscopy (SEM) images of *Allium* seed coat patterns. We used the following techniques: K-means, K-means++, Minibatch K-means, Spectral, and Balanced Iterative Reducing and Clustering utilizing Hierarchies (Birch). Because of their ability to discover similarities and differences in data, they are an excellent solution for exploratory data analysis. We present how to form groups using unsupervised machine learning.

1. We proposed to classify seed coat patterns using unsupervised machine learning.
2. We then compared them to previous human-based classifications.
3. Following that, we suggested our proposed classification based on unsupervised machine learning and possible combinations such as SI, SS, SM, NS, PS, and PD.

2. Results

In this section, we report two types of results comparisons: a comparison between the results of human- and machine-based classification methods, and a comparison between the results of our combinations. This allows us to determine the relationship between each method. We selected three types of threshold values for data scales. First, we selected a minimum threshold value of 500 because we believed that it could provide useful information and was a reasonable value. We then selected a medium value between 1000 and 2000, and a maximum value above 2000. By increasing the threshold, we strongly expected that the results would be the same.

2.1. Comparison between Results of Human-Based and Machine-Based Methods

We first report the results for each relationship in tables provided for the machine-based cluster. As briefly explained in Figure 1, we named the relationship of each group. That is, (1) method name, (2) cluster name, and (3) separation number. The method and cluster are present in this case, but since one method depends on another, it must be separated again. If the threads M1C1-3 and M3C1 match M3C1 and M4C3 in Figure 2a, it means that M1C1-3 is connected to M1C1 three times. In addition, M1C1 and M2C1-2 match with M2C1-2 and M4C3 in Figure 2b. Figure 2d shows the correspondence between M1C2 and M3C1 and M2C2 and M3C3. Therefore, this relationship can be considered as a cluster image of comparable threads.

Figure 2b shows that the numerical value for identical M2C3-1 and M3C1 and M2C3-2 and M3C2 ranges from 1120 to 1127. Moreover, the numerical value for identical models, M2C1 and M2C1-1, M2C2 and M3C1-2, M2C3 and M2C1-1, M2C4 and M3C2-2, M2C5 and M3C3, M2C5 and M3C6, M4C1 and M5C4-1, M4C5 and M5C1-2, M4C6 and M5C1-2′s in Figure 2c ranges from 1120 to 1127. Therefore, these images can be considered identical. However, there is no relationship in Figure 2c,d. In Figure 2d, only a relation of 1312 is shown.

Related M2C3 and M3C3 in one thread that corresponds to M3C1 and M2C3 in Figure 3c. In both groups Figure 3a,b, there is no connection and no thread.

For the thread, five-colored clusters of ST, associated with M1C2 and M3C1 correspond to M2C2 and M3C3 in Figure 4d. Therefore, this model provides reasonable support for clustering and very similar clusters.

For identical: M2C2 and M3C1-1, M2C2-3 and M3C2-1, M2C3-1 and M4C6-1, M2C5-1 and M4C4, M2C5-2 and M4C5, M2C2 and M3C1-2, M2C1 and M5C5, M2C2 and M3C1-2, M2C2-4 and M5C2, and M2C3-2 and M4C2 in Figure 4b. Moreover, the values of M2C2-2 and M3C1-1, M2C2-3 and M3C2-1, M2C3-2 and M4C6-2, M2C5-1 and M4C4, M2C5-2 and M4C5, M2C2-1 and M5C5, M2C2-4 and M5C2, and M3C2-2 and M4C2 are between 911 and 912, which means that they are very similar images. For Figure 4a,c, there is no relation and similar values, only 4d shows at least one relation.

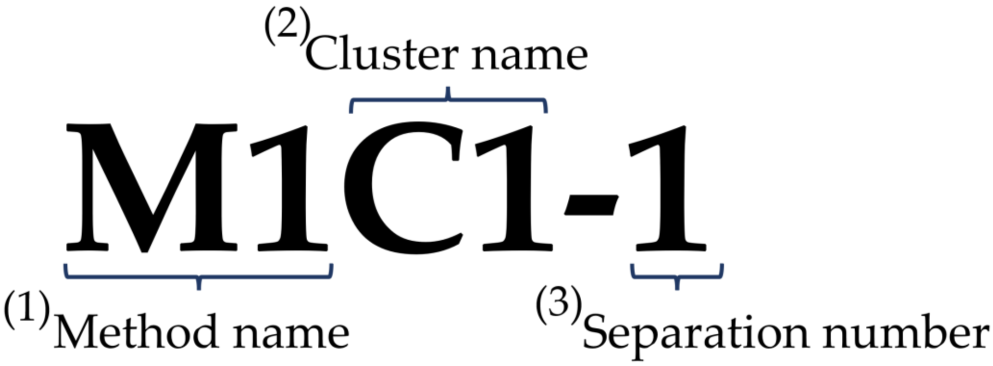


Figure 1. Example relationship of the group for separations.

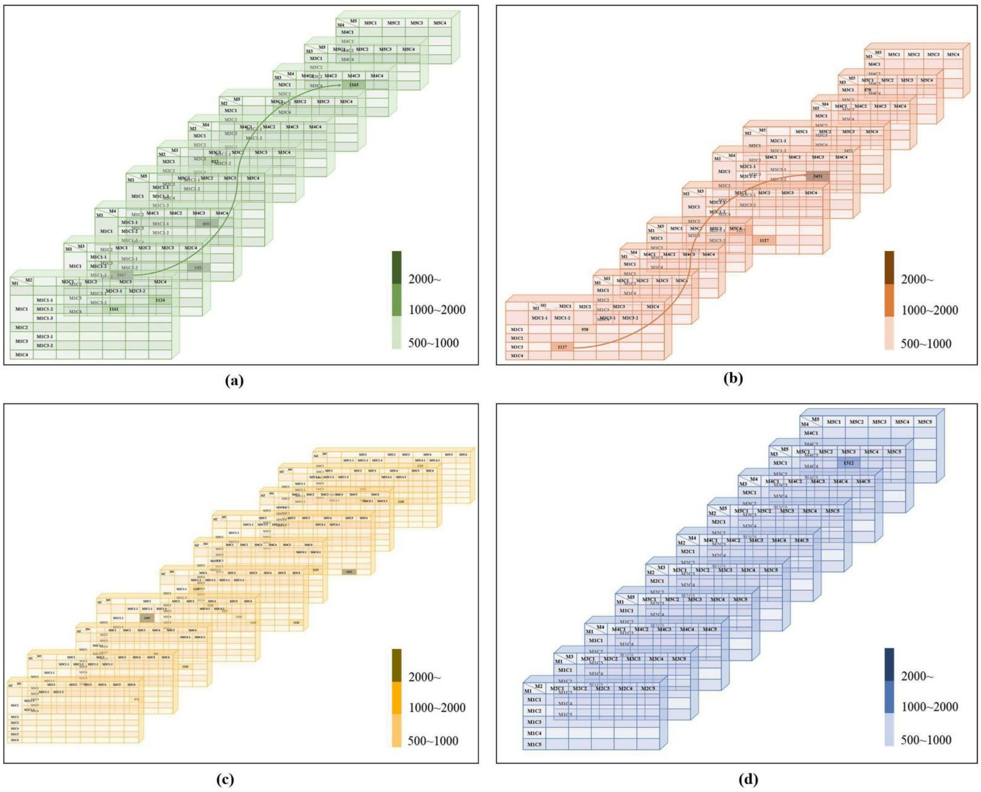


Figure 2. Anticlinal wall result–straight results. (**a**) four-grayscale clusters result, (**b**) four-threshold clusters result, (**c**) six-grayscale clusters result, (**d**) five-colored clusters result.

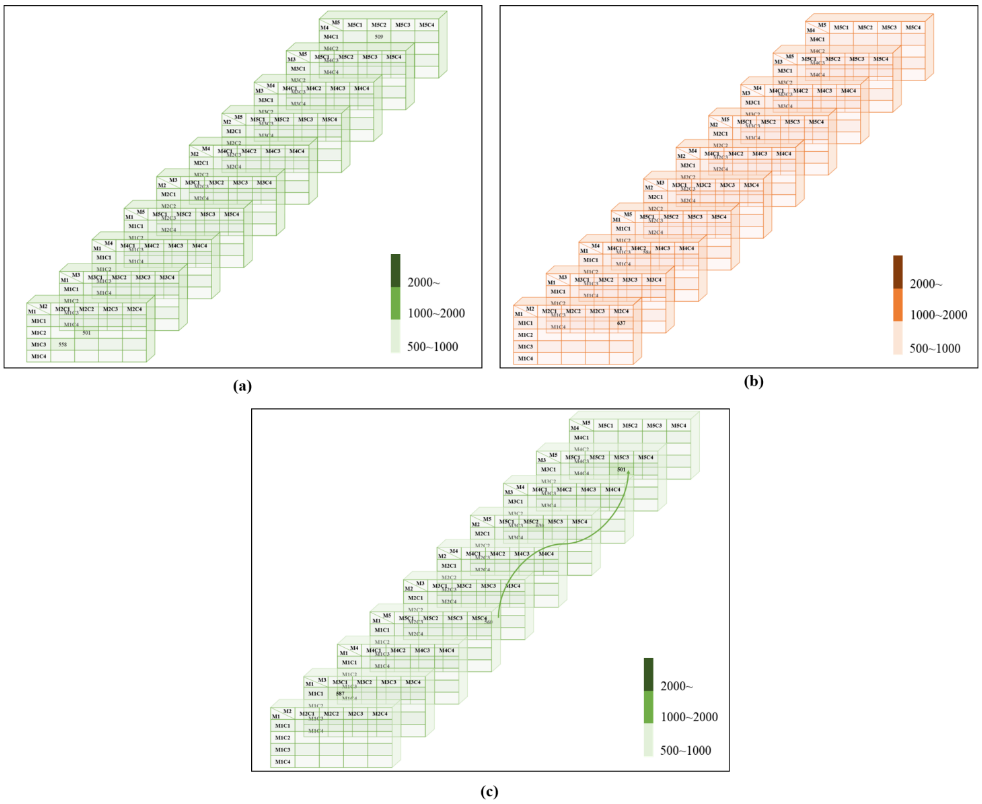
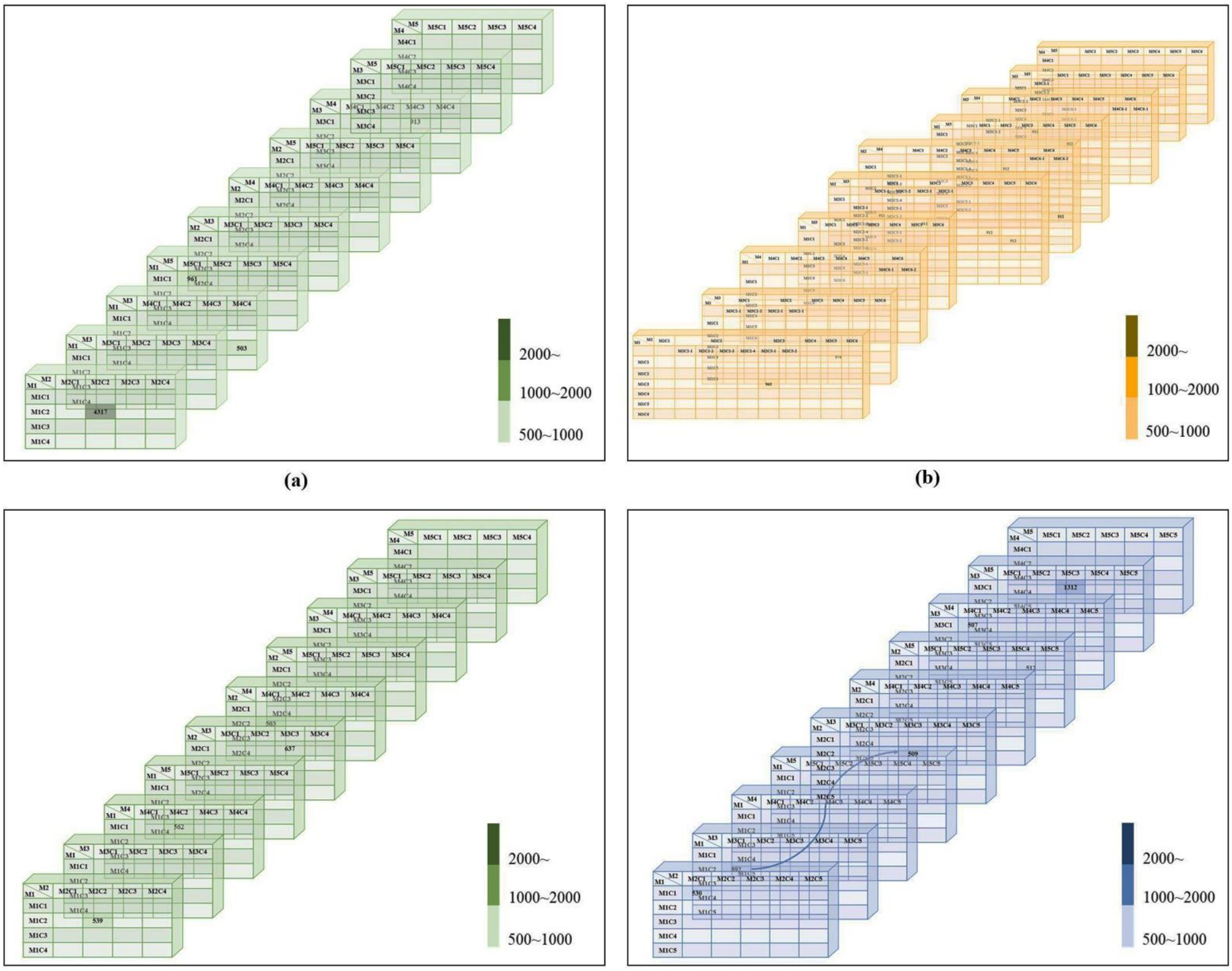


Figure 3. Anticlinal wall results. U to Ω—(**a**) four-grayscale clusters result, (**b**) four-threshold clusters result, U—(**c**) four-grayscale clusters result.



(a) (b) (c) (d)

Figure 4. Periclinal wall results. Marginal verrucae—(**a**) four-grayscale clusters result, (**b**) six-grayscale clusters result, Small verrucae—(**c**) four-grayscale clusters result, (**d**) five-colored clusters result.

2.2. Proposed Clustering

In this section, we present the main clustering generated by a machine. For this purpose, we used grayscale, threshold, and colored image clustering through the elbow and silhouette methods. Using these methods, we compare the number of cluster images that can represent a relation and a thread between these types of techniques and determine the number of images that can be clustered together. Table 1 presents the abbreviations and shows how we define them in this paper. Table 2 presents the main clustering of machine-based observations. In addition, the data in Table 2 belongs to Figures 5 and 6.

Table 1. Proposed picking grouping.

Our Observations	Abbreviation	Definition
Strongly identical	SI	It refers to an image that is strongly similar.
Strongly similar	SS	It is referring term refers to an image that is exactly similar.
Similar	SM	It refers to one or more images that are similar to one another.
Nearly similar	NS	It indicates that two or more images are similar to one another.
Possibly similar	PS	It indicates that there is a chance that certain images are identical.
Possibly different	PD	It is indicating that there is a chance that some images may be different.

Table 2. Related image number for each group.

SI	SS	SM	NS	PS	PD
1385	1127	1127	845	824	869

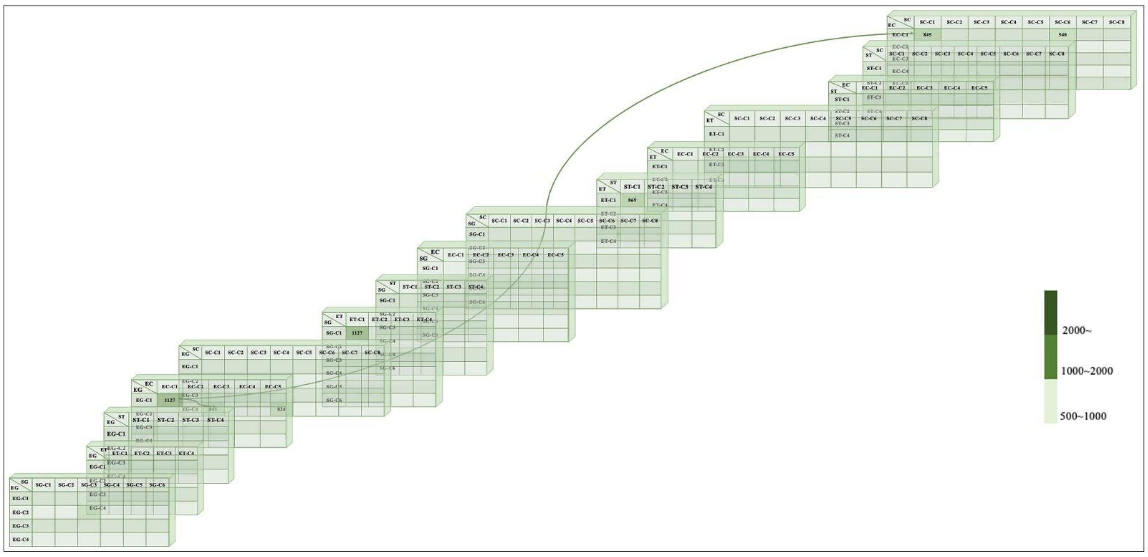
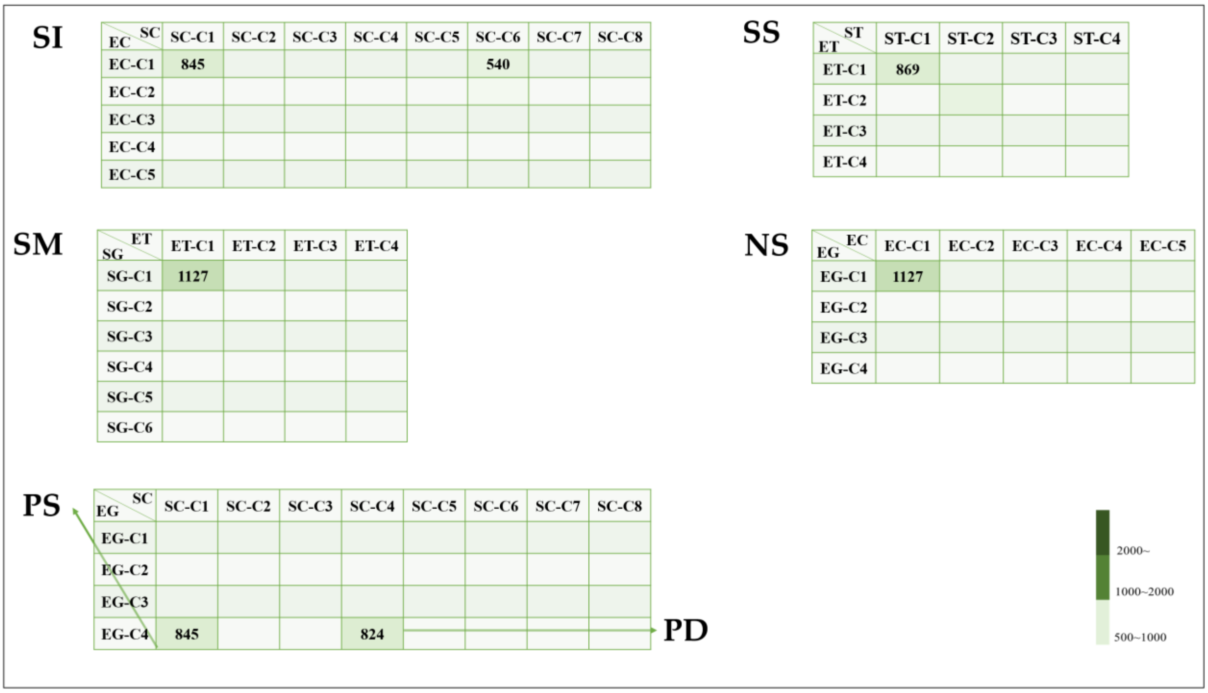


Figure 5. Proposed clustering.


SI

EC SC	SC-C1	SC-C2	SC-C3	SC-C4	SC-C5	SC-C6	SC-C7	SC-C8
EC-C1	845					540		
EC-C2								
EC-C3								
EC-C4								
EC-C5								

SS

ET ST	ST-C1	ST-C2	ST-C3	ST-C4
ET-C1	869			
ET-C2				
ET-C3				
ET-C4				

SM

SG ET	ET-C1	ET-C2	ET-C3	ET-C4
SG-C1	1127			
SG-C2				
SG-C3				
SG-C4				
SG-C5				
SG-C6				

NS

EG EC	EC-C1	EC-C2	EC-C3	EC-C4	EC-C5
EG-C1	1127				
EG-C2					
EG-C3					
EG-C4					

PS

EG SC	SC-C1	SC-C2	SC-C3	SC-C4	SC-C5	SC-C6	SC-C7	SC-C8
EG-C1								
EG-C2								
EG-C3								
EG-C4	845			824				

PD
2000~
1000~2000
500~1000

Figure 6. Detailed information on the proposed clustering.

Figure 5 illustrates the model of the proposed clustering. As for the threads, related EG-C1 and EC-C1 correspond to EG-C4 and SC-C1 correspond to EC-C1 and SC-C1. This approach seems to suggest that these clusters are similar threads. Figure 6 shows detailed information about each group. Therefore, this relation is more than one of anticlinal and periclinal wall threads.

For the identical approach, the numerical value of EG-C4 and SC-C1 and EC-C1 and SC-C1 is 845 and that of EG-C1 and EC-C1 and SG-C1 and ET-C1 is 1127. These values are equal to the anticlinal straight result of the six-grayscale clusters and four-threshold clusters.

Therefore, we have mentioned the most important observation regarding machine-based clustering. In this case, where there is more than one thread of human-based clustering, we can see that the thread, EG, SC-C1, and EC-C1′2 images may be the same. In these cases, the information given by the two thread approaches was similar due to the similarity of the information from the threads.

Next, we determine the type of image used. For Figure 7, the five most prevalent images were selected from each image in the six-method testing. This allows us to determine the group the images came from. The result is that the images from 7a to 7d are similar, the images from 7e are slightly distinct, and the images from 7f are largely different from the others.

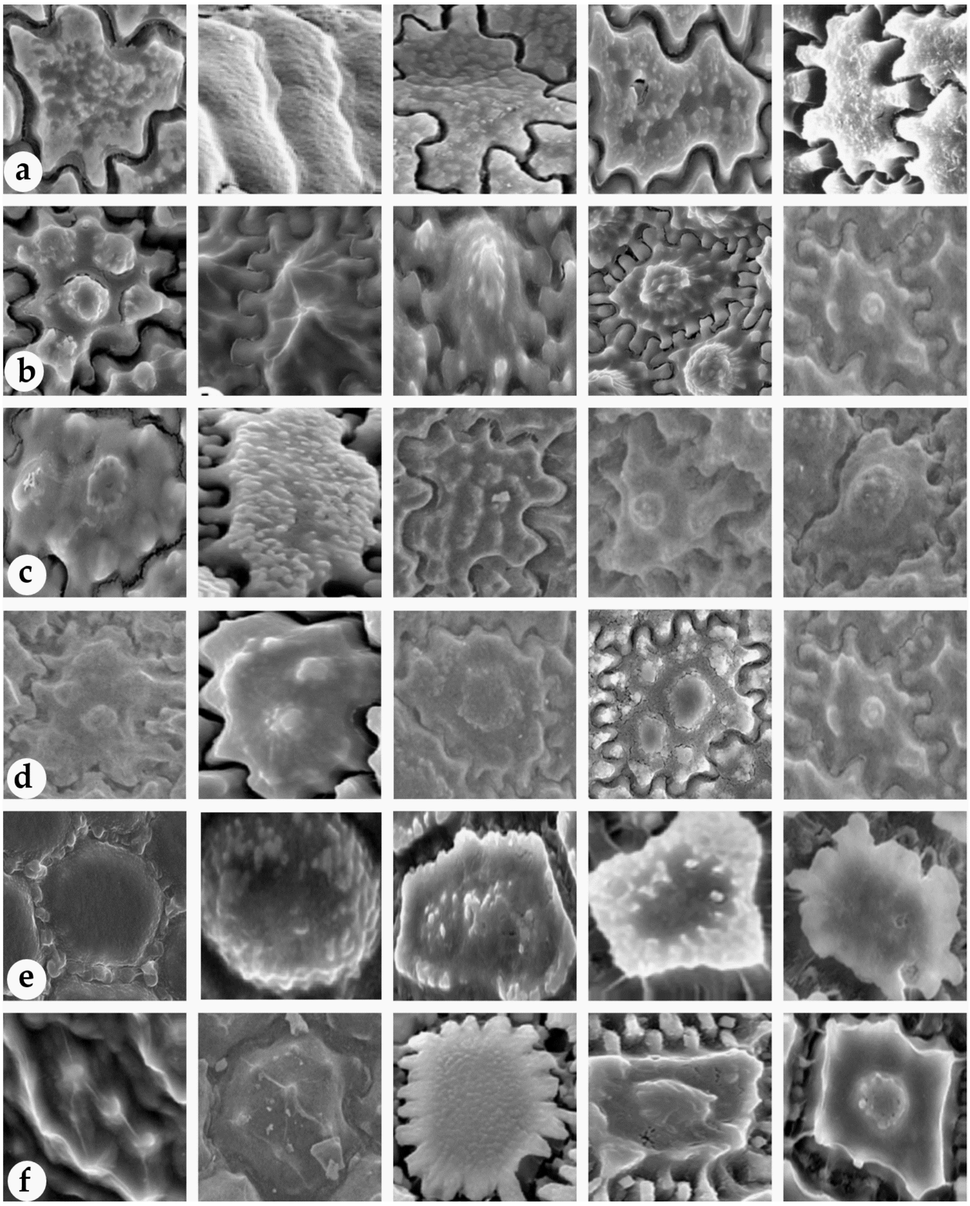


Figure 7. Related images for each group from [5]. (**a**) Actual images in SI, (**b**) Identical image in SS, (**c**) Similar image in SM, (**d**) Approximately similar image in NS, (**e**) Possibly similar image in PS, (**f**) Remaining images in PD.

3. Discussion

Many researchers have studied the deep- and machine-learning-based classification of plant organs such as plant disease classification [25,26], and tomato-type classification [27]. However, there is no research on the identification and classification of seed coat patterns using machine learning or deep learning to date. The seed testa sculpture is one of the most important taxonomic features for plants. Many researchers determined seed coat shape patterns, sculpturing of the periclinal wall, and type of undulation of anticlinal walls to identify species of *Allium* [1,2,5–16]. However, humans identify different characteristics of seed coat sculptures. Therefore, we identified the seed coat pattern of *Allium* images using deep learning for the first time.

In our approach, the use of unsupervised machine learning methods has provided the opportunity to investigate new types of morphology, including ultrastructure. We attempted to determine our own clustering, which provides another proposal using unsupervised machine-learning techniques for the clustering of *Allium* seeds. We presented the proposed analysis by showing the quantities of each group in Figure 5 and the images that best fit our criteria in Figures 6 and 7.

We made comparisons between all models and our model. In the highlighted analysis, we show the kind of images included in each case. We obtained five common images from each group. In the SI case, 540 images were from the EC-C1 and SC-C6 group, whereas 845 images were from the EC-C1 and SC-C1. The SS case contained 869 images from the SG-C1 and ET-C1 groups. As for the SM case, it contained 1127 images from SG-C1 and ET-C1. The EG-C1 and EC-C1 images were taken from 1127 images in the NS group. In addition, the EG-C1 and SC-C1 groups contributed 1127 images. Another 845 images were taken by the EG-C4 and SC-C1 in the PS group. The PD group also included 824 images from EG-C4 and SC-C4.

The second part of our analysis compared our final group with the five methods used. Because the five-method group was obtained using human-based clustering it showed overlapping anticlinal for straight, U, and U to Ω, and periclinal wall types for marginal verrucae and small verrucae. Indeed, the two relations shown in Figure 6 are proposed approaches. In contrast, human-based clustering showed one relation in Figure 2a,b, Figures 3c and 4d. The reason we mention the previous association is that we want to highlight our own clustering for a distinct concept other than conventional classification. Using the two prior cases, we then also identified a similar case. The value of small verrucae in the four-threshold group's count value for the human-based group value ranging from 1120 to 1127. In addition, the count value of the marginal verrucae six-grayscale case was between 911 and 912. The count value for EG-C4 and SC-C1 and EG-C1 and SC-C1 was between 1120 and 1127 and that for EG-C1 and EC-C1 and SG-C1 and ET-C1 was 845. The identical count value for the final group was the same as for the first five method groups. We discovered SI, SS, SM, NS, PS, and PD using the proposed approach. The SI group outperformed the SS, SM, NS, PS, and PD groups. In addition, the SS, SM, NS, and PS groups are very similar. Moreover, the PD groups are still distinct.

4. Materials and Methods

4.1. The Proposed Module Overviews

The proposed algorithm was implemented by grouping the data in an unlabeled dataset based on the underlying hidden features in the data in Figure 8, which consisted of two phases: (a) general module and (b) performance analysis module. These two phases are explained in more detail in the following subsections.

4.2. General Module

4.2.1. Datasets

We selected the seed coats of over 100 *Allium* species from previous studies [5,8–10,30]. The seed coat sculptures of *Allium* species are displayed in Figure 9. The species names along with descriptions of anticlinal and periclinal walls as well as corresponding references of all selected species are given in Supplementary Material.

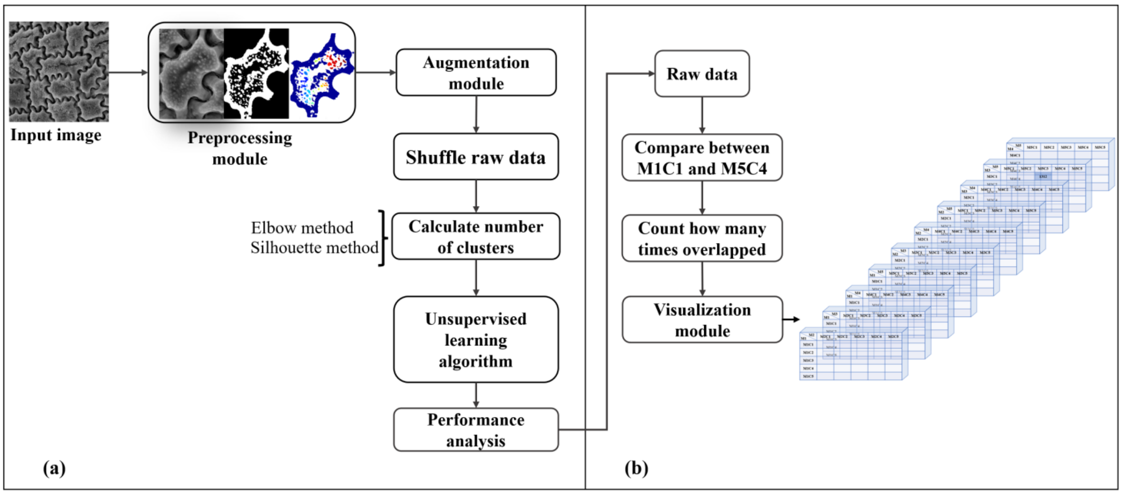


Figure 8. Proposed algorithm for this study. (**a**) General module, (**b**) Performance analysis module.

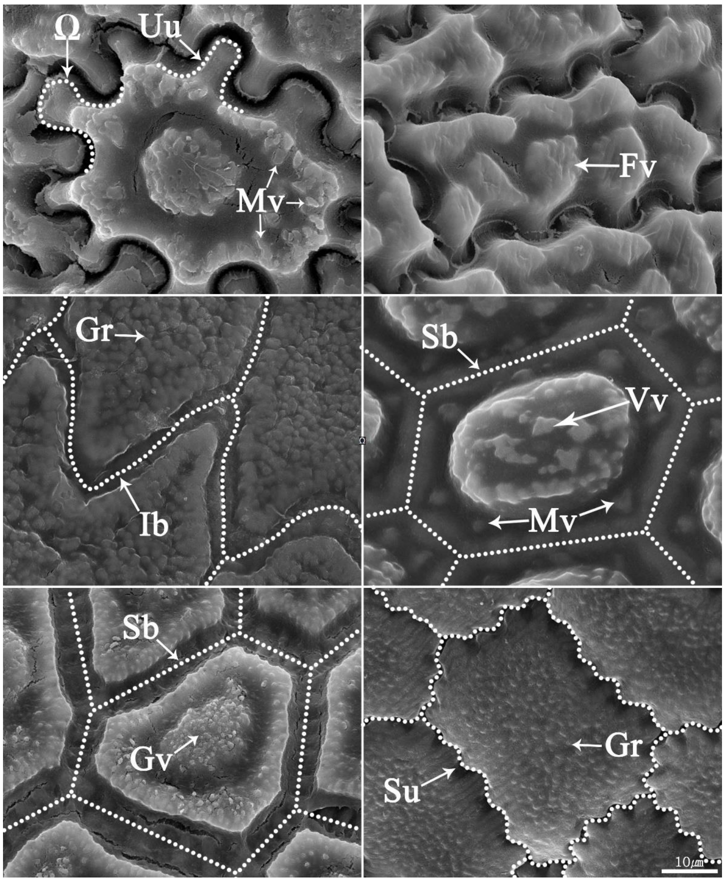


Figure 9. Type of seed testa of *Allium* derived from [5]. Ω, Omega-type undulation; Uu, U-type undulation; Su, S-type undulation; Ib, irregular boundary; Sb, straight boundary; Gr, granule; Vv, verrucate verruca; Mv, marginal verruca; Gv, granulate verruca.

Figure 9 depicts seed coat types for anticlinal and periclinal walls. Anticlinal walls are the boundaries between two perpendicular cell walls, whereas periclinal walls are the cells outside the surface [31]. Among the seeds depicted in the picture, we selected three and two characters from the anticlinal and periclinal walls, respectively. The most overlapping types in the test results were straight, U, U to Ω from the anticlinal wall and marginal verruca, small verruca from the periclinal wall.

4.2.2. Preprocessing Module

To prepare the image for the main system operation (the image in Figure 8), preprocessing is required to deal with varied input data sizes, noise, and color. First, the goal is to crop the seed walls separately (Figure 10a–f). This ensures dimensional uniformity in all images used in the next step. Each image is normalized at this point, by cropping it to 250 × 250 pixels, to subsample it into the desired dimension. Preprocessing involves images such as grayscale images (Figure 10i), threshold images (Figure 10k), and colored images (Figure 10l). For the first preprocessing of the grayscale image, the input image was converted to a grayscale image. Grayscale representations are often used to extract descriptors rather than color images directly because grayscales reduce computational costs and simplify algorithms [32,33]. Color might be of limited interest in many applications, and the introduction of unnecessary data might increase the required training data. In the following preprocessing stage, the pixel values were divided into white backgrounds or black foregrounds using a threshold [34].

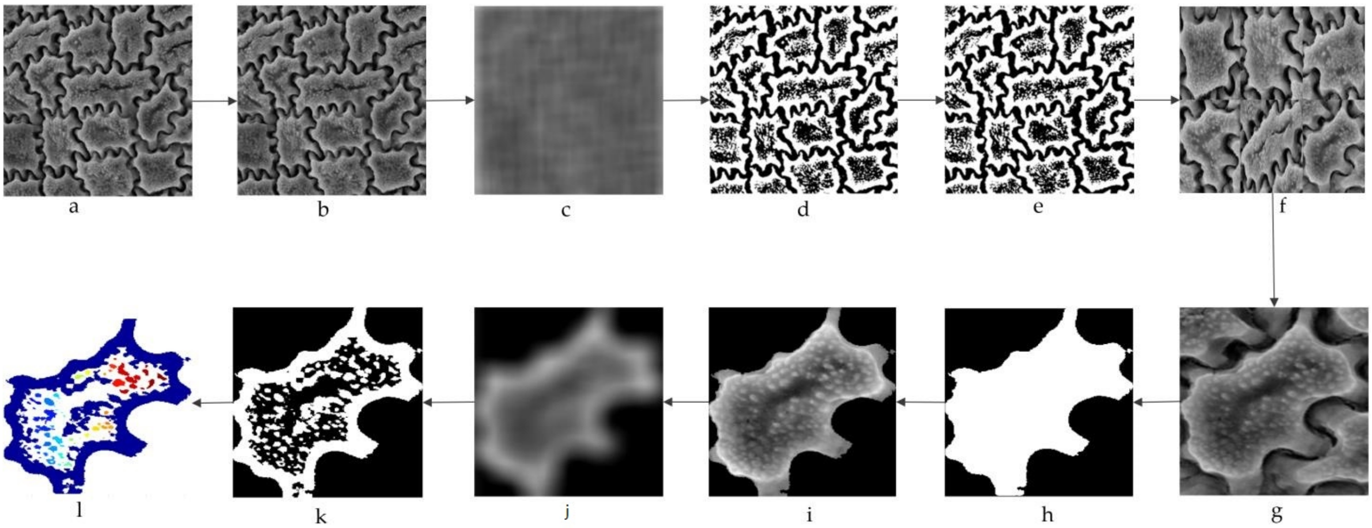


Figure 10. Data preprocessing of the seed coat pattern. (**a**) Input image, (**b**) grayscale image, (**c**) threshold using adaptive thresholding, (**d**) binary image, (**e**) labeled image (**f**) and cropped image, (**g**) separated cropped image, (**h**) masked image, (**i**) grayscale image, (**j**) binary image obtained using an adaptive threshold, (**k**) binary image, (**l**) and colored image.

Thresholding is the simplest image segmentation method and the most common way to convert a grayscale image to a binary image. In the next step, we selected a threshold value, and all the gray level values smaller than the threshold value were classified as 0 (black, or background) while all the gray level values equal to or greater than the threshold value were classified as 1 (white, or foreground). It is important to remember that grayscale images contain pixels with values between 0 and 1, therefore the threshold value is within a closed range [0.0, 1.0] [35,36]. In this case, we set the threshold to 0.6. The pixel larger than 0.6 indicates an original pattern in white, and the remaining pixels indicate a background pattern. In the final preprocessing of the colored image, we applied the label-connected component to the threshold image. The label-connected component associated with the label has the constraint of accepting a binary image [37,38]. It can also support a label matrix which is usually output from a label component connected to the label as well as binary images. This binary image should contain a set of objects that are separated from each other. Pixels that belong to objects are labeled 1 and pixels that are in the background are labeled 0. In fact, an object is those pixels that are 1 that are connected in a chain when considering local neighborhoods [37]. It gives the affiliation of each pixel. This indicates where each pixel belongs if it falls on an object. In the final preprocessing phase, the applied measurement measures the different image quantities and features in a black-and-white image. More specifically, based on black and white, it automatically determines the properties of each contiguous white region that is eight-connected. The purpose of the final stage was to visualize the color assigned to each object based on the number of objects in the label matrix. Once the preprocessing phase is complete, the images are immediately passed on to the next process.

4.2.3. Augmentation Module

To increase the diversity of the dataset, a suitable data augmentation technique is needed to improve the size and quality of the training set by creating modified data from the existing data [39,40]. The purpose of this technique is to extend and improve the dataset to reflect deformations in diverse seed patterns in a real-world setting. In our study, random rotation was performed with a maximum angle of 5°, width and height shifts were performed with a value of 0.01, and shifting pixels from left to right and top to bottom was performed.

4.2.4. Shuffling Raw Data

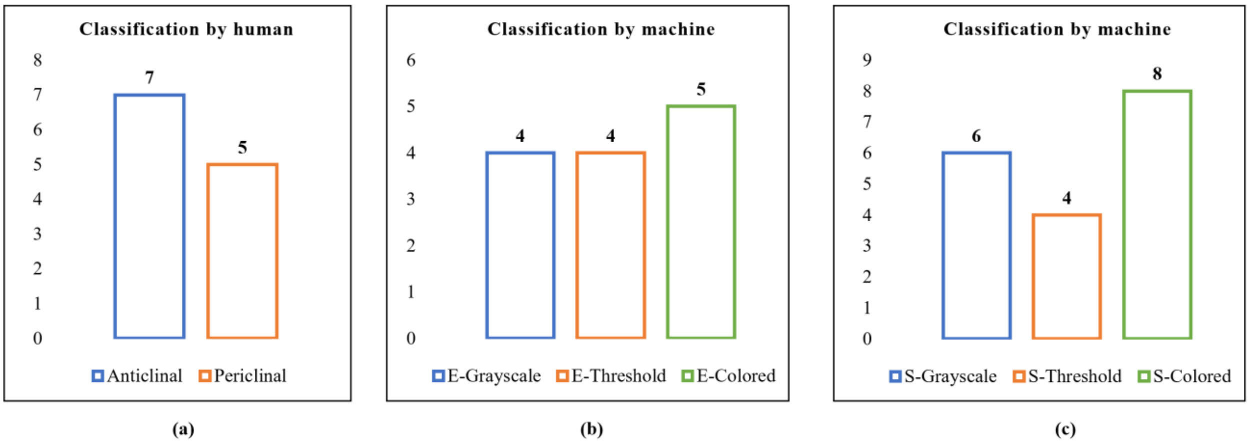
We then split the data into separate training and testing sets. Before training the machine learning model, it is important to thoroughly shuffle the dataset to avoid bias or patterns within the split dataset [41]. The goal of dataset shuffling is to improve the quality and predictive performance of machine-learning models. In this study, we randomly shuffled the training dataset.

4.2.5. Calculating the Number of Clusters

One of the biggest challenges in unsupervised learning is determining the number of clusters needed. Therefore, it is necessary to determine the number of clusters in advance of clustering [42,43]. We applied the elbow and silhouette schemes, which are useful techniques in evaluating the quality of clustering, to determine the optimal number of clusters in our study. The elbow method is easy to implement because the k number of clusters is based on the sum of squared distance (SSE) between the data points and their associated cluster centroids [44–46]. In the silhouette technique, unlike the elbow method, the silhouette coefficient is calculated and the number of clusters is easily determined [46]. The value of the silhouette coefficient ranges from $[-1, 1]$, where a high score indicates that the sample is far from the neighboring clusters and a low score indicates that these samples may have been assigned to the wrong cluster [47]. Thus, we applied the elbow and silhouette schemes to determine the optimal number of clusters, changing the value between 2 and 10. Figure 11a shows the original cluster created by human-based clustering. In contrast, Figure 11b shows the number of clusters obtained by the elbow method, and Figure 11c shows the number of clusters obtained by the silhouette method.

4.2.6. Clustering Module

The goal of unsupervised methods is to determine whether a group of data is formed based on similarities between individual pieces of information. Consequently, cluster analysis is an excellent method to study the relationships between groups. In this study, we used K-means, K-means++, Minibatch K-means, Spectral, and Birch clustering, which are extremely easy and among the most popular unsupervised machine-learning methods. It may be difficult and arbitrary for a machine to decide the number of clusters to form in K-means, one of the most commonly used algorithms for cluster analysis [48], when the number of clusters is given in advance. For clustering analysis, we used K-means clustering to compute cluster indices, centroid locations, and distances between points and centroid positions. Next, we introduced a popular variant of the classic K-means algorithm, named K-means++. Bahmani et al. [49] demonstrated how this algorithm can significantly improve the quality of clustering by improving the initialization of centroids. Therefore, we also used K-means++ and included widely separated centroids. This increases the likelihood of initially picking up centroids that are in different clusters, and since the centroids are picked up from the data points, each centroid ends up with some data points associated with it. There is a version of the K-means algorithm in unsupervised machine learning known as Minibatch K-means. The algorithm creates random batches of data for storage in memory and then collects a random batch of data at each iteration to update the cluster [50]. It also facilitates cluster search by reducing computational costs. It is also a widely used technique in spectral cluster analysis. Spectral clustering uses a connectivity approach for clustering where communities of connected or immediately adjacent nodes are identified graphically [51]. The nodes are then mapped into a low-dimensional space that can be easily separated to form clusters. The Birch clustering algorithm can cluster a large dataset by first creating a small and compact summary of the large dataset that contains as much information as possible [52]. It uses hierarchical methods to cluster and reduce data. Birch only needs to scan the data set in a single pass to perform clustering [52].



(a) (b) (c)

Figure 11. (**a**) Original number of clusters generated by humans. (**b**) Number of clusters generated using the Elbow method. (**c**) Number of clusters generated using the Silhouette method.

4.3. Performance Analysis

In this section, we show a bi-directional representation of the community using our proposed performance analysis. We used five types of unsupervised clustering analyses to characterize overlapping relationships. We first introduce how to evaluate the performance analysis results of data groups or clusters, which consist of two types of analyses: the comparison between human- and machine-based analysis, and our main clustering. Table 3 shows the abbreviations used in our study.

Table 3. Proposed unsupervised techniques in our study.

Techniques (1)	Abbreviation (1)	Clusters (2)	Abbreviation (2)	Proposed Cluster (3)	Abbreviation (3)
K-Means	M1	Cluster 1	C1	Elbow based four-grayscale	EG
K-Means++	M2	Cluster 2	C2	Silhouette based six-grayscale	SG
Minibatch K-means	M3	Cluster 3	C3	Elbow based four-threshold	ET
Spectral	M4	Cluster 4	C4	Silhouette based four- threshold	ST
Birch	M5	Cluster 5	C5	Elbow based five-colored	EC
				Silhouette based eight- colored	SC

We first illustrate the combinations algorithm for overlapping relation needed in this study in Figure 12.

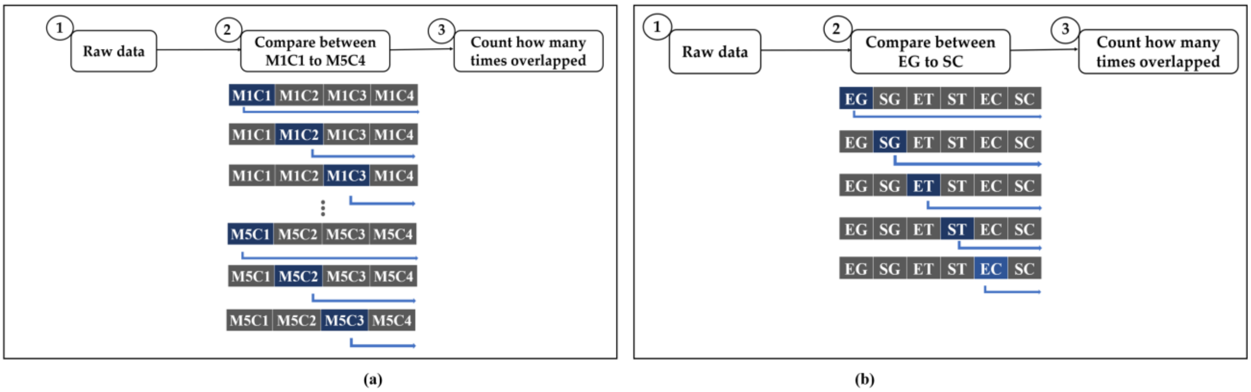


Figure 12. Overlapping algorithm overview. Each method generates overlap relations as well as threads based on the results. (**a**) Comparison between human-based and machine-based algorithm (**b**) Our proposed algorithm.

Figure 12a shows a comparison between the methods used to analyze relationships. In this section, we explain how we use this algorithm. First, we read the test image of the raw data for performance analysis. Then, we compare the results of the methods we find related. For example, it would be useful to compare M1C1 and M1C2, M1C1 and M1C3, and M1C1 and M1C4 as shown in Figure 12a. This process will be completed after M5C4. Thereafter, the number of times overlapping values occurred is recorded. Figure 12b illustrates the main analysis. It is the same process as before. However, the comparison procedure is different from that represented in Figure 12a. The difference is that we used the elbow and silhouette method to determine the number of clusters required. We found several types of clusters. Therefore, we compared EG and SC, EG and ET, EG and ST, EG and EC, EG and SC.

Visualization Module

After completing the performance analysis, we introduce the visualization module, which we used to connect the entire result of cube-based visualization. In the cube-based visualization, the number of cubes we need is obtained from Equation (1). Equation (2), then gives the amount of the tables. In the formula, k represents the number of possibilities for the selected objects, ! represents factorial, and n represents the total number of samples in the set.

Mathematically, the formula for determining the number of arrangements by selecting only a few objects from a set without repetition can be expressed as follows:

$$C_n^k = \frac{n!}{k!(n-k)!} \tag{1}$$

Therefore, the total number of tables based on the preceding methodology is as follows:

$$\sum tables \tag{2}$$

5. Conclusions

This study used unsupervised machine learning to categorize the seed coat patterns of *Allium* species. We then compared the categories with previous human-based classifications. Following that, we proposed and examined unsupervised machine learning and possible combinations, which are SS, SI, SM, NS, PS, and PD. The anticlinal wall type of straight obtained the best results, with four-grayscale clusters and four-threshold clusters, while the anticlinal wall type of U to Ω produces good results with four-grayscale clusters. Furthermore, the periclinal wall of small verrucae five-colored cluster produces the best result. There was one anticlinal and periclinal wall relationship that was strongly related. The proposed SI group outperformed the SS, SM, NS, PS, and PD groups, whereas the SS, SM, and NS outscored the PD group. Therefore, unsupervised machine-learning algorithms were discovered to be suitable for grouping seed coat patterns.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/plants11223097/s1, Table S1: The species names along with descriptions of anticlinal and periclinal walls as well as corresponding references of all selected species.

Author Contributions: Research direction and experimental design: S.C.; performing the experiments: G.A.; research discussion: S.C., G.A., S.B., J.C.L.K. and H.J.C.; manuscript preparation: G.A. and S.B.; manuscript review: S.C., H.J.C. and J.C.L.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Korea Meteorological Administration Research and Development Program under Grant KMI 2021-01310.

Institutional Review Board Statement: Not Applicable.

Informed Consent Statement: Not Applicable.

Data Availability Statement: All data generated or analyzed during this study are included in this published article.

Acknowledgments: All authors appreciate the support of the Changwon National University. This study is a part of the M.S. thesis of the first author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Friesen, N.; Fritch, R.M.; Blattner, F.R. Phylogeny and new intrageneric classification of *Allium* (Alliaceae) based on nuclear ribosomal DNA ITS sequences. *Aliso* **2006**, *22*, 372–395. [CrossRef]
2. Choi, H.J.; Oh, B.U. A partial revision of *Allium* (Amaryllidaceae) in Korea and northeastern China. *Bot. J. Linn. Soc.* **2011**, *167*, 153–211. [CrossRef]
3. Xie, D.F.; Tan, J.B.; Yu, Y.; Gui, L.J.; Su, D.M.; Zhou, S.D.; He, X.J. Insights into phylogeny, age and evolution of *Allium* (Amaryllidaceae) based on the whole plastome sequences. *Ann. Bot.* **2020**, *125*, 1039–1055. [CrossRef]
4. POWO 2022. Plants of the World Online. Facilitated by the Royal Botanic Gardens, Kew. Available online: http://www.plantsoftheworldonline.org/ (accessed on 10 October 2022).
5. Baasanmunkh, S.; Lee, J.K.; Jang, J.E.; Park, M.S.; Friesen, N.; Chung, S.; Choi, H.J. Seed morphology of *Allium* L. (Amaryllidaceae) from Central Asian countries and its taxonomic implications. *Plants* **2020**, *9*, 1239. [CrossRef]
6. Yusupov, Z.; Ergashov, I.; Volis, S.; Makhmudjanov, D.; Dekhkonov, D.; Khassanov, F.; Tojibaev, K.; Deng, T.; Sun, H. Seed macro- and micromorphology in *Allium* (Amaryllidaceae) and its phylogenetic significance. *Ann. Bot.* **2022**, *129*, 869–911. [CrossRef]
7. Baasanmunkh, S.; Choi, H.J.; Oyuntsetseg, B.; Friesen, N. Seed testa sculpture of species of *Allium* L. (Amaryllidaceae) and its taxonomic implications. *Turczaninowia* **2021**, *24*, 154–161. [CrossRef]
8. Choi, H.J.; Giussani, L.M.; Jang, C.G.; Oh, B.U.; Cota-Santez, J.H. Systematics of disjunct northeastern Asian and northern North American *Allium* (Amaryllidaceae). *Botany* **2012**, *90*, 491–508. [CrossRef]
9. Celep, F.; Koyuncu, M.; Fritsch, R.M.; Kahraman, A.; Dogan, M. Taxonomic importance of seed morphology in *Allium* (Amaryllidaceae). *Syst. Bot.* **2012**, *37*, 893–912. [CrossRef]
10. Lin, C.Y.; Tan, D.Y. Seed testa micromorphology of thirty-eight species of *Allium* (Amaryllidaceae) from central Asia, and its taxonomic implications. *Nord. J. Bot.* **2017**, *35*, 189–200. [CrossRef]
11. Barthlott, W. Epidermal and seed surface characters of plants: Systematic applicability and some evolutionary aspects. *Nord. J. Bot.* **1981**, *1*, 345–355. [CrossRef]
12. Barthlott, W. Scanning electron microscopy of the epidermal surface in plants. In *Scanning Electron Microscopy in Taxonomy and Function Morphology*; Claugher, D., Ed.; Clarendon Press: Oxford, UK, 1990; Volume 41, pp. 69–94.
13. Kruse, J. Rasterelektronenmikroskopische Untersuchungen an Samen der Gattung *Allium* L. *Die Kult.* **1984**, *32*, 89–101. [CrossRef]
14. Kruse, J. Rasterelektronenmikroskopische Untersuchungen an Samen der Gattung *Allium* L. II. *Die Kult.* **1986**, *34*, 207–228. [CrossRef]
15. Kruse, J. Rasterelektronenmikroskopische Untersuchungen an Samen der Gattung *Allium* L. III. *Die Kult.* **1988**, *36*, 355–368. [CrossRef]
16. Kruse, J. Rasterelektronenmikroskopische Untersuchungen an Samen der Gattung *Allium* L. IV. *Feddes Repert.* **1994**, *105*, 457–471. [CrossRef]
17. Geetharamani, G.; Pandian, A. Identification of plant leaf diseases using a nine-layer deep convolutional neural network. *Comput. Elect. Eng.* **2019**, *76*, 323–338. [CrossRef]

18. Lee, S.H.; Chan, C.S.; Mayo, S.J.; Remagnino, P. How deep learning extracts and learns leaf features for plant classification. *Pattern Recognit.* **2017**, *71*, 1–13. [CrossRef]
19. Reyes, A.K.; Caicedo, J.C.; Camargo, J.E. Fine-tuning Deep Convolutional Networks for Plant Recognition. *CLEF* **2015**, *1391*, 467–475.
20. Saleem, M.H.; Potgieter, J.; Arif, K.M. Plant disease classification: A comparative evaluation of convolutional neural networks and deep learning optimizers. *Plants* **2020**, *9*, 1319. [CrossRef]
21. Wang, X.; Liu, J.; Zhu, X. Early real-time detection algorithm of tomato diseases and pests in the natural environment. *Plant Methods* **2021**, *17*, 1–17. [CrossRef]
22. Meunkaewjinda, A.; Kumsawat, P.; Attakitmongcol, K.; Srikaew, A. Grape leaf disease detection from color imagery using hybrid intelligent system. In Proceedings of the 2008 5th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, Krabi, Thailand, 14–17 May 2008; Volume 1, pp. 513–516. [CrossRef]
23. Padol, P.B.; Yadav, A.A. SVM classifier based grape leaf disease detection. In Proceedings of the 2016 Conference on Advances in Signal Processing (CASP), Pune, India, 9–11 June 2016; pp. 175–179. [CrossRef]
24. Sannakki, S.S.; Rajpurohit, V.S.; Nargund, V.B.; Kulkarni, P. Diagnosis and classification of grape leaf diseases using neural networks. In Proceedings of the 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), Tiruchengode, India, 4–6 July 2013; pp. 1–5. [CrossRef]
25. Mahlein, A.K. Plant disease detection by imaging sensors–parallels and specific demands for precision agriculture and plant phenotyping. *Plant Dis.* **2016**, *100*, 241–251. [CrossRef]
26. Atila, Ü.; Uçar, M.; Akyol, K.; Uçar, E. Plant leaf disease classification using EfficientNet deep learning model. *Ecol. Inform.* **2021**, *61*, 101182. [CrossRef]
27. Brahimi, M.; Boukhalfa, K.; Moussaoui, A. Deep learning for tomato diseases: Classification and symptoms visualization. *Appl. Artif. Intell.* **2017**, *31*, 299–315. [CrossRef]
28. Amara, J.; Bouaziz, B.; Algergawy, A. A deep learning-based approach for banana leaf diseases classification. *BTW* **2017**, *266*, 79–88.
29. Piazza, G.; Valsecchi, C.; Sottocornola, G. Deep Learning Applied to SEM Images for Supporting Marine Coralline Algae Classification. *Diversity* **2021**, *13*, 640. [CrossRef]
30. Veiskarami, G.; Khodayari, H.; Heubl, G.; Zarre, S. Seed surface ultrastructure as an efficient tool for species delimitation in the *Allium ampeloprasum* L. alliance (Amaryllidaceae, Allioideae). *Microsc. Res. Technol.* **2018**, *81*, 1275–1285. [CrossRef]
31. Koch, K.; Bhushan, B.; Barthlott, W. Multifunctional surface structures of plants: An inspiration for biomimetics. *Prog. Mater. Sci.* **2009**, *54*, 137–178. [CrossRef]
32. Saravanan, C. Color image to grayscale image conversion. In Proceedings of the Second International Conference on Computer Engineering and Applications (ICCEA), Bali, Indonesia, 19–21 March 2010; Volume 2, pp. 196–199. [CrossRef]
33. Kanan, C.; Cottrell, G.W. Color-to-grayscale: Does the method matter in image recognition? *PLoS ONE* **2012**, *7*, e29740. [CrossRef]
34. Singh, T.R.; Roy, S.; Singh, O.I.; Sinam, T.; Singh, K. A new local adaptive thresholding technique in binarization. *Int. J. Comput. Sci.* **2012**, *8*, 271–277. [CrossRef]
35. Al-amri, S.S.; Kalyankar, N.V.; Khamitkar, S.D. Image Segmentation by Using Threshold Techniques. *arXiv* **2010**, arXiv:1005.4020.
36. He, J.; Do, Q.D.M.; Downton, A.C.; Kim, J. A comparison of binarization methods for historical archive documents. *Proc. Int. Conf. Doc. Anal. Recognit.* **2005**, *1*, 538–542. [CrossRef]
37. Di Stefano, L.; Bulgarelli, A. A simple and efficient connected components labeling algorithm. In Proceedings of the International Conference on Image Analysis and Processing, Venice, Italy, 27–29 September 1999; pp. 322–327. [CrossRef]
38. He, L.; Chao, Y.; Suzuki, K.; Wu, K. Fast connected-component labeling. *Pattern Recognit.* **2009**, *42*, 1977–1987. [CrossRef]
39. Shorten, C.; Khoshgoftaar, T.M. A survey on Image Data Augmentation for Deep Learning. *J. Big. Data* **2019**, *6*, 1–48. [CrossRef]
40. Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; Yang, Y. Random erasing data augmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 13001–13008. [CrossRef]
41. Elmahdy, A.; Mohajer, S. On the fundamental limits of coded data shuffling for distributed machine learning. *IEEE Trans. Inf. Theory* **2020**, *66*, 3098–3131. [CrossRef]
42. Masud, M.A.; Huang, J.Z.; Wei, C.; Wang, J.; Khan, I.; Zhong, M. I-nice: A new approach for identifying the number of clusters and initial cluster centres. *Inf. Sci.* **2018**, *466*, 129–151. [CrossRef]
43. Kim, S.W.; Gil, J.M. Research paper classification systems based on TF-IDF and LDA schemes. *Hum. Cent. Comput. Inf. Sci.* **2019**, *9*, 1–21. [CrossRef]
44. Bholowalia, P.; Kumar, A. EBK-means: A clustering technique based on elbow method and k-means in WSN. *Int. J. Comput. Appl.* **2014**, *105*, 17–24. [CrossRef]
45. Syakur, M.A.; Khotimah, B.K.; Rochman, E.M.S.; Satoto, B.D. Integration k-means clustering method and elbow method for identification of the best customer profile cluster. *Mater. Sci. Eng.* **2018**, *336*, 012017. [CrossRef]
46. García, S.; Parejo, A.; Personal, E.; Guerrero, J.I.; Biscarri, F.; León, C. A retrospective analysis of the impact of the COVID-19 restrictions on energy consumption at a disaggregated level. *Appl. Energy* **2021**, *287*, 116547. [CrossRef]
47. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Comput. Appl. Math.* **1987**, *20*, 53–65. [CrossRef]
48. Sinaga, K.P.; Yang, M.S. Unsupervised K-means clustering algorithm. *IEEE Access* **2020**, *8*, 80716–80727. [CrossRef]

49. Bahmani, B.; Moseley, B.; Vattani, A.; Kumar, R.; Vassilvitskii, S. Scalable K-Means++. *Proc. VLDB Endow* **2012**, *5*, 622–633. [CrossRef]
50. Sculley, D. Web-scale k-means clustering. In Proceedings of the 19th International Conference on World Wide Web, Raleigh, NC, USA, 26–30 April 2010; pp. 1177–1178. [CrossRef]
51. Aggarwal, C.; Subbian, K. Evolutionary network analysis: A survey. *ACM Comput. Surv.* **2014**, *47*, 1–36. [CrossRef]
52. Zhang, T.; Ramakrishnan, R.; Livny, M. BIRCH: An efficient data clustering method for very large databases. *SIGMOD Rec.* **1996**, *25*, 103–114. [CrossRef]