Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Application of Unsupervised Clustering Techniques to Somatic Embryos for Automated SE Fluidics System

Chaitanya S. Nayak
Georgia Institute of Technology
Punnag Chatterjee
Indian Institute of Technology
Cyrus Aidun
cyrus.aidun@me.gatech.edu

Georgia Institute of Technology

Research Article

Keywords: Clustering, Embryo Morphology, SE Fluidics System, Somatic Embryogenesis, Unsupervised Learning

Posted Date: June 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9589135/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Application of Unsupervised Clustering techniques to Somatic Embryos for automated SE Fluidics system

Chaitanya S. Nayak[1,2], Punnag Chatterjee[3], Cyrus Aidun[1,2*]

[1]Renewable Bioproducts Institute, Georgia Institute of Technology, Atlanta, 30332, Georgia, USA.
[2]George W. Woodruff School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, 30332, Georgia, USA.
[3]Mechanical, Materials, Aerospace Engineering, Indian Institute of Technology, Dharwad, 580011, Karnataka, India.

*Corresponding author(s). E-mail(s): cyrus.aidun@me.gatech.edu;
Contributing authors: cnayak6@gatech.edu;
punnag.chatterjee@iitdh.ac.in;

Abstract

Somatic embryogenesis (SE) is a viable method for clonal propagation of economically significant crops including forest trees such as pines and spruces. However, large-scale SE production requires automation to reduce manual labor, especially in selecting and harvesting mature embryos. Since embryo maturation is asynchronous, identifying viable embryos remains a challenge. This study explores machine learning techniques to cluster grayscale images of mature SEs processed through a fluidics-based SE system for image analysis and sorting. A dataset of 720 grayscale embryo images with a subset of embryos labelled as 'accepted' or 'rejected' based on the embryo morphology was selected for this study. The full dataset was clustered using a range of UL methods. The K-means method, which requires human intervention to select the k value, with Principal Component Analysis (PCA) achieved 82% accuracy for rejected and 92% for acceptable embryos. Constrained Unsupervised Learning (CUL), where pairwise constraints were introduced to nudge the UL method to improve the clustering accuracy of segregating clusters, compared to morphology-based clustering achieved nearly 100% accuracy for acceptable embryos and 87.5% for rejected embryos. Using an unconstrained hierarchical clustering technique, dendrograms were plotted and analyzed to understand cluster formations at different threshold values for

new image datasets that consisted of different numbers of grayscale images (500, 3,500, and 4,089). It was found that the dataset consisting of 500 images sufficiently represents the exhaustive morphological traits of the embryos. The results show that a small number of embryos in the range of few hundred can be used to provide the base cluster that can be used for identifying embryos with acceptable morphology. Lastly, a supervised learning (SL) model was trained on the clusters extracted from CUL, achieving 100% segregation of rejected embryos. The trained SL model was 32% faster than morphological methods. The proposed pipeline ensures efficient in-situ embryo categorization, facilitating scalable automation in industrial SE applications while minimizing human bias.

Keywords: Clustering, Embryo Morphology, SE Fluidics System, Somatic Embryogenesis, Unsupervised Learning

1 Introduction

The transition to a carbon-neutral future heavily depends on sustainable forest management [1] and the use of improved genetic materials [2]. Forests act as vital carbon sinks, offsetting emissions through their growth and storage capabilities. Planted forests, in particular, play a significant role in enhancing this carbon sequestration potential [3]. Additionally, wood biomass offers a sustainable energy source with higher energy efficiency compared to other plant-based materials [4]. European forests have been a significant carbon sink for the past few decades, but variations exist across regions due to differences in forest area, structure, and harvesting intensity. These variations are influenced by natural forest expansion, afforestation, and improved management practices, all of which contribute to enhanced carbon sequestration and storage in different parts of Europe [5]. Among tree species, conifers have traditionally dominated forestry operations, particularly in the northern hemisphere, and are gaining prominence in the southern hemisphere. Norway spruce (Picea abies), a cornerstone species in European forestry, holds immense economic and ecological value [6]. It dominates European conifer forests, making it a crucial species for forestry and ecological studies. Focusing on its embryos ensures a comprehensive understanding of its developmental traits, aiding in improved propagation techniques. This is essential for conservation, commercial forestry, and advancing somatic embryogenesis research. Also, this species has as investigated, great carbon sequestration potential for climate mitigation across Europe, with sequestration rates reaching up to 500 g C m^{-2} yr^{-1}, even in high-latitude regions with short growing seasons [7]

Recent advancements have demonstrated the feasibility of automating SE plant production processes via SE fluidics systems [8], paving the way for large-scale applications in sustainable forestry. A key aspect of automating somatic embryogenesis (SE) is the ability to identify and characterize morphological traits that correlate with germination-competent, mature embryos [9]. While Machine Learning (ML) has seen extensive application in agriculture, including tissue culture, its potential for SE remains underexplored. In the agriculture industry, image processing was introduced almost a decade ago to aid the sorting and classification of seeds for crop production

[10] [11]. [12] captured images to measure the physical properties of embryogenic callus using artificial neural networks (ANN). Genetic algorithms have been used to predict and optimize the constituents in the culture medium used in tissue culture studies [13] [14]. However, there needs to be more research exploring the viability of using ML algorithms as an alternative to morphological feature-based sorting algorithms. For this purpose, image datasets of somatic embryos were, obtained using the SE Fluidics system [8] [15]

Using the images captured with the SE Fluidics system, [16] recently explored the classification of embryo grayscale images using supervised learning (SL) techniques such as convolutional neural networks (CNN) and support vector machines (SVM). The work provided valuable insights into the preprocessing steps and in-line classification of images into predefined categories such as "Accept Left," "Accept Right," and "Reject" based on its prominent characteristics. SL methods inherently involve human intervention in defining classes for obtaining labeled datasets on which the SL model is trained. Researchers usually decide the number of classes based on the visually prominent features. This manual process often relies on superficial features to classify, potentially overlooking significant morphological characteristics within the image dataset. Determining the optimal number of classes to represent the required unique features of the dataset remains challenging. This paper takes the next step by investigating the automation in the segregation of embryo image datasets, using unsupervised learning (UL) techniques (including K-means, hierarchical clustering, and constrained or semi-supervised algorithms). These approaches extend the framework established by Chatterjee et al. (2024), addressing the challenges and advancing the insights into the segregation process.

2 Methods: Technique Development & Results

The SE fluidic system records the images when embryos are detected to flow within a square glass tube, illuminated with an LED light source, and backed by a black-colored paper. The images are captured using a Prosilica GC1290 camera fitted with a 25mm fixed focal length lens from Edmund Optics Inc. The images were saved in a monochrome format with an image size of 501 $\times$ 400 pixels and approximately 70 pixels/mm resolution (15-μm pixel size). The raw images obtained from the SE fluidics imaging system are pre-processed to obtain grayscale images ready for further analysis and clustering. Transformation and processing of images are shown in Fig. 1 , The pre-processing begins by cropping the grayscale $I(x, y)$, defined as a 2D surface function, to keep only the region containing the embryo. This is done by identifying a rectangular area, the bounding box $B = (x_{\min}, y_{\min}, x_{\max}, y_{\max})$ around the embryo and discarding unnecessary portions. The cropped image is then resized to a fixed dimension (M, N) using bilinear interpolation, where new pixel values are estimated based on the surrounding pixels' intensities to maintain smooth transitions. In order to eliminate unwanted elements from an image, a detection step finds the regions of interest $T(x, y)$, which are subsequently replaced out using a smoothing method like Gaussian inpainting. Optical Character Recognition (OCR) was used to identify the unwanted elements in our situation, which were texts. However, similar methods can

be used to eliminate other kinds of noise, such as scratches, artifacts, or unwanted objects. The following equation shows application of the process to remove them:

$$I_{\text{no-text}}(x, y) = G_\sigma \, I(x, y)$$

, where G_σ represents a Gaussian filter that smooths and blends the removed $T(x, y)$, into the surrounding image to maintain consistency. Object masking follows, where a threshold

$$T_{\text{adaptive}}(x, y) = \mu(x, y) + C$$

is determined dynamically based on the local mean intensity $\mu(x, y)$ and manually fine-tuning the constant C. This method ensures robust separation of the embryo from the background, even under varying lighting conditions. In our case, the lighting condition for all the images was kept constant. After segmentation, dilation, erosion and structuring, we refined the embryo's shape by eliminating noise and smoothing its boundaries. This is done using a closing operation

$$I_{\text{refined}} = (I_{\text{binary}} \oplus S) \ominus S$$

, where $\oplus$ (dilation) fills small gaps in the embryo, and $\ominus$ (erosion) removes small noise artefacts, with S as the structuring element, usually a unity matrix, defining the operation's size and shape. Finally, all non-embryo regions are set to a uniform black background using $I_{\text{final}}(x, y)$, where pixel values are preserved for the embryo and set to zero elsewhere. This structured workflow ensures the removal of unwanted artefacts while preserving the embryo's morphological details for further analysis.

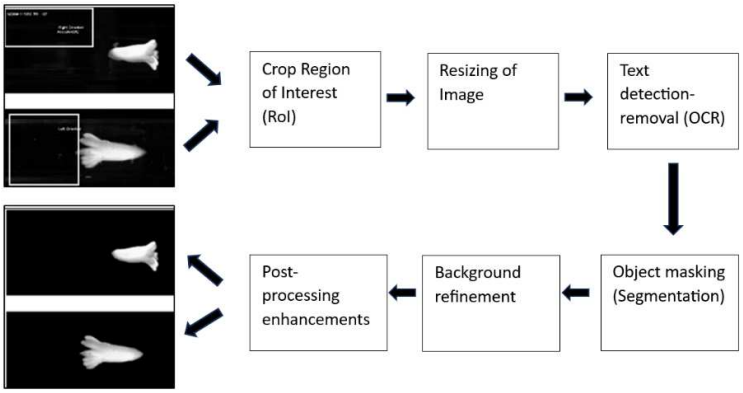


Fig. 1 Raw embryo images with spots, smudges and texts, captured using SE Fluidics system, (pairs shown on the top left); Corresponding processed and ready to use (for clustering) images (pairs shown on the bottom left).

2.1 Ensuring morphology-based classification

This study dealt with image dataset consisting of 720 grayscale images of embryos captured using the SE Fluidics system. These images vary in perspective and orientation

due to differences in embryo positioning during capture. As a result, the orientation of an embryo in an image may differ depending on factors such as cotyledon alignment (left or right) and the angle of the body length relative to the horizontal and vertical axes of the image. Despite this variability, the goal is not to cluster based on orientation. Instead, the dataset is clustered based on morphological features, as shown in Fig. 2, including body length, cotyledon count, cotyledon spacing, and the size and shape of cotyledons and the embryo body.

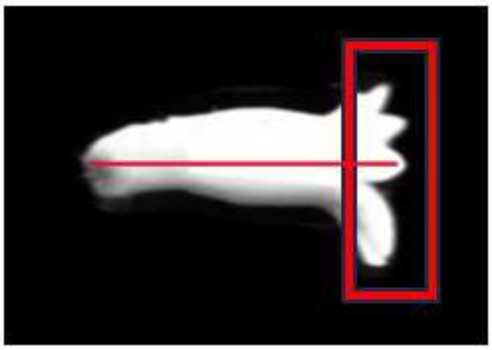


Fig. 2 Body length of a Norway spruce embryo spanning from the body tip until the root as shown by red horizontal line; five (one behind) distinct cotyledons demarcated within the red box.

In SE fluidics systems, embryos are imaged from two different perspectives, top-view and side-view, to capture their three-dimensional structure. Although these images appear different due to perspective changes, they remain dependent on each other because they represent the same embryo at the same developmental stage. This means that key biological features, such as the number of cotyledons (embryonic leaves), their arrangement, the overall shape of the embryo, and the spacing between cotyledons, remain the same regardless of the viewing angle. While the way light interacts with the embryo and the shape distortion due to perspective may alter pixel values in the image, the underlying structure does not change. In this study, to maintain consistency in clustering, only one standardized perspective, side-view, was used for all embryos, ensuring that each embryo was represented by a single image. To simulate different viewing angles without introducing a second independent perspective, image augmentation techniques were applied. In the first stage, it included rotating each image by 0°, 90°, 180°, and 270° and applying flips along both the horizontal and vertical axes, creating multiple versions of the same image. The purpose of this augmentation was to prevent the clustering algorithm from grouping images solely based on orientation rather than meaningful features. These meaningful features, referred to as intrinsic morphological traits, include measurable aspects of the embryo's physical form, such as its overall size, cotyledon count, body length, and cotyledon spacing, characteristics that define its development and classification. By using a single perspective while applying controlled transformations, the study ensured that clustering was based on these developmentally relevant properties rather than accidental differences in image orientation.

In the second stage, Histogram of Oriented Gradients (HOG) vectors were extracted for every image variation. Sangeetha and Deepa (2017) [17] analytically describes the HOG feature extraction process, detailing the conversion of an image into its corresponding feature vector by computing gradient magnitudes and orientations, followed by spatially local histogram binning and normalization. This approach enhances scale invariance and robustness, ensuring that extracted features remain consistent across variations in embryo morphology, thereby improving clustering and classification accuracy.

These vectors were averaged to form a single representative HOG vector for each image, capturing attributes across all variations. This averaged vector represented a new data point in a higher-dimensional space determined by the dimensions of the HOG vectors at a particular setting for cell, bin, and block sizes. The averaging process created a centroid for each image's variations, effectively reducing orientation-based clustering and emphasizing morphological features. The third stage involved analyzing the variance within each cluster to evaluate the distribution of data points. A low variance indicated that points were tightly grouped, suggesting that clustering was influenced primarily by orientation. In contrast, a high variance suggested that clustering was driven by broader morphological differences rather than orientation alone. However, even in high-variance clusters, data points remained grouped due to shared dominant morphological characteristics, such as cotyledon count and spacing. This variance analysis helped distinguish whether clustering was governed by subtle orientation changes or more pronounced morphological features. Once clustering was confirmed based on morphology, baseline methods involving pure unsupervised learning (K-means, Unconstrained Hierarchical) and Semi-Supervised Learning techniques were implemented, detailed in the next section.

2.2 User-based clustering: Application of K-means clustering algorithm

This section examines the effectiveness of K-means in performing clustering without external supervision, in contrast to SSL, which operates under predefined constraints. The investigation evaluates how these fully unsupervised techniques identify patterns within grayscale embryo images based solely on inherent feature distribution. By transitioning to an unconstrained clustering approach, it provides insights into the strengths and limitations of constrained and unconstrained clustering, to help identify which method better aligns with the research objectives. K-means clustering is an unsupervised machine learning algorithm that partitions a dataset into a predefined number of clusters, where the parameter k serves as a constraint on the number of clusters. The objective is to ensure that data points within the same cluster exhibit greater similarity to each other than to those in different clusters [18]. When applied to 720 grayscale embryo images, the method involved preprocessing steps such as Histogram of Oriented Gradients (HOG) feature extraction and Principal Component Analysis (PCA) for dimensionality reduction.

By setting k=3, the goal was to cluster images into three clusters (akin to human defined Accept Left, Accept Right, and Reject categories). The clustering produced three groups, as shown in Fig. 3. The first cluster primarily contained 148 Accept

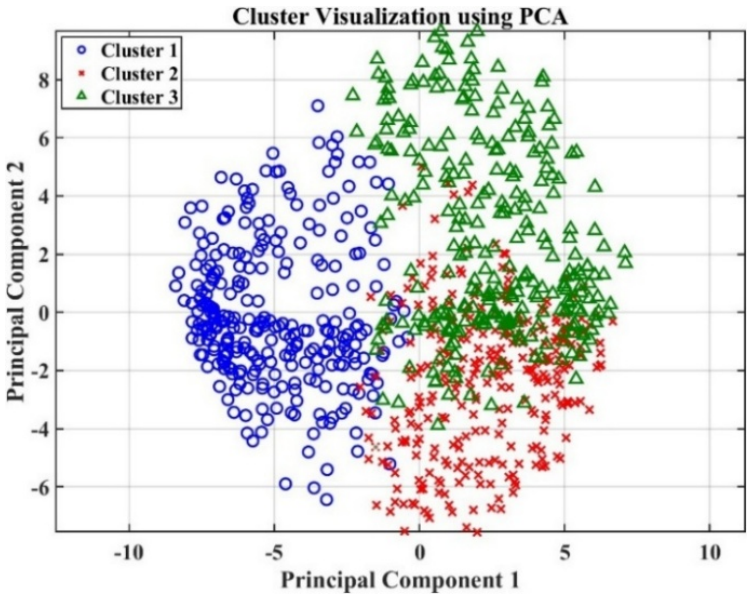


Fig. 3 PCA plot used to visualize K-means clustering technique for k=3. The 720 embryo images are segregated into three clusters; Cluster 1 (blue-circle: reject), Cluster 2 (red-cross: accept left); Cluster 3 (green-triangle: accept right).

Left images with some overlap from other categories, the second had 162 Accept Right images with minor mixing, and the third predominantly consisted of 410 Reject images with minimal miscategorisation. These results highlighted both the strengths and limitations of K-Means in separating categories based on pixel intensity similarities. Although K-Means clustering somewhat aligned with the desired classifications, accuracy remained low at 47.6%, 53.2%, and 83.9% for Accept Left, Accept Right, and Reject categories, respectively, suggesting challenges in capturing features when forced to use k=3.

Testing k=2 simplified grouping and yielded improved results. The first cluster contained 207 Reject images with minor overlap, while the second included 513 Accept images with minimal miscategorisation, as shown in Fig. 4, achieving 82% accuracy for the Reject cluster and 92% accuracy for the merged Accept cluster, which combines the previously separate Accept-Left and Accept-Right clusters. This demonstrates that reducing k allowed the algorithm to focus on the primary categories, effectively separating Accept and Reject embryos by analyzing morphological traits and disregarding irrelevant factors such as orientation.

2.3 Semi-Supervised Learning: Constrained UL with pairwise constraints

Semi-Supervised Learning (SSL) enhances clustering by leveraging limited labelled data, such as must-link and cannot-link pairs, alongside a larger set of unlabeled data. Lu and Leen [18] introduced an optimization-based formulation that modifies

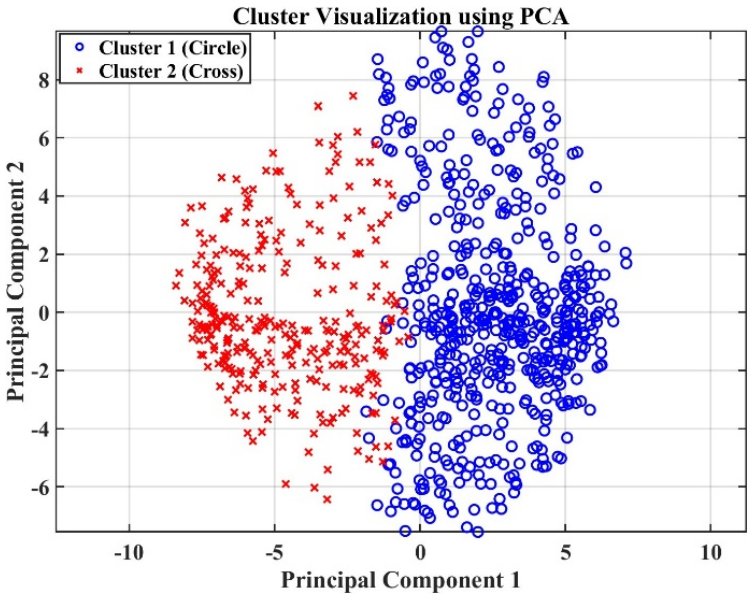


Fig. 4 PCA plot used to visualize K-means clustering technique for k=2. 720 embryo images segregated into two clusters as just discussed; Cluster 1 (blue-circle: reject-left), Cluster 2 (red-cross: accept-right).

the clustering objective function to incorporate these constraints, thereby improving clustering accuracy in scenarios where full supervision is impractical. Their approach minimizes the following objective function:

$$J = \sum_{i=1}^{n} \sum_{k=1}^{K} h_{ik} \left\| x_i - \mu_k \right\|^2 + \lambda \sum_{(i,j) \in M} h_{ik} h_{jk} - \lambda \sum_{(i,j) \in C} h_{ik}(1 - h_{jk})$$

Here, x_i represents an individual data point (an embryo image in this context), and μ_k denotes the center (or mean position) of cluster k. The first term in the equation measures how far each data point is from its assigned cluster center, ensuring that similar points are grouped closely. The second term enforces must-link constraints, meaning that if two points i and j belong to the same must-link set M, they should be assigned to the same cluster. The third term enforces cannot-link constraints, ensuring that points in the cannot-link set C remain in different clusters. The weight factor λ determines how strongly the must-link and cannot-link constraints influence the clustering outcome. The variable h_{ik} is a binary indicator, taking a value of 1 if point x_i belongs to cluster k, and 0 otherwise. This formulation enables clustering based on intrinsic embryo features like cotyledon count and shape while reducing orientation-based variations. It enhances unsupervised classification by providing structural guidance without extensive manual labelling. Feature extraction techniques, such as contour detection and spatial analysis, further refine biologically meaningful clusters.



Fig. 5 (a) On the left, the pair of "Well-Defined" embryos; (b) The middle pair is the pair of "Elongated Blobs"; (c) The rightmost pair is of "Circular Blobs.

Their work provides a mathematical foundation for semi-supervised clustering, but to use it on images, it must be adapted into an algorithm that can process image features. Hence, to determine pairwise constraints and generate them for embryo images, we refer to Masud et al. [19] who derive must-link and cannot-link constraints based on local density and similarity measures. This enables automated constraint generation, refining clustering without manual annotation. Masud et al. (2019) proposed a method to generate pairwise constraints for semi-supervised clustering by analyzing the density and neighborhood structure of an unlabeled dataset. Given a dataset $X = \{x_1, x_2, x_3, \ldots, x_n\}$ (in our case $X = 720$ images), similarity between data points is computed using a distance metric, such as the Euclidean distance, defined as:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$$

, where x_i & x_j are feature vectors in an m-dimensional space of i^{th} and j^{th} images. Local density ρ_i at each point is estimated based on the number of neighbors within a user predefined radius ε, given by:

$$\rho_i = \sum_j \Theta(\varepsilon - d(x_i, x_j))$$

, where Θ is the Heaviside function. Must-link constraints are assigned to points in high-density regions, satisfying $d(x_i, x_j) < \varepsilon$, ensuring that strongly similar points remain in the same cluster. Conversely, cannot-link constraints are applied to points with significantly different densities or those separated by $d(x_i, x_j) > \lambda\varepsilon$, enforcing separation between dissimilar data. These constraints refine clustering by guiding the process toward meaningful groupings (via human intervention) while preserving the data-driven nature of unsupervised learning. This approach is particularly relevant to the objectives set, as it enables automatic derivation of pairwise constraints from an unlabeled embryo dataset, improving clustering accuracy without requiring manual annotations. These pairs of images served as inputs to direct clustering towards morphological features, such as body length and cotyledon count. Using this approach, embryos were categorized into three distinct groups as shown in Fig. 5. The first category included well-defined embryos, Fig. 5 (a) with a distinct body length and three or more cotyledons. The second group comprised elongated blobs, Fig. 5 (b) embryos with a positive body length but indistinguishable or absent cotyledons. The final category included circular blobs, Fig. 5 (c) characterized by embryos lacking both body length and cotyledons. The clustering process involved grouping 720 grayscale embryo images into three categories using pairwise constraints. Must-link constraints paired

embryos with similar morphological traits, such as clustering well-defined embryos with other well-defined embryos or circular blobs with other circular blobs. Cannot-link constraints paired one well-defined embryo with a circular blob or an elongated blob to ensure clear separation. The clustering results showed that the well-defined cluster contained 454 images with a 100% accuracy rate when compared to human morphology-based classification, grouping all embryos with three or more cotyledons. The circular blobs cluster included 144 images with 87.5% accuracy, comprising 126 circular blobs and 18 elongated blobs without cotyledons. Lastly, the elongated blobs cluster had 122 images with 100% accuracy, grouping embryos with elongated bodies but no cotyledons. Pairwise constraints significantly influenced the clustering process, aligning the results with the intended morphological categories. The semi-supervised learning algorithm was tested by dividing the dataset into two clusters based on the number of cotyledons in each embryo. The intended classification was to separate embryos with three or more cotyledons into one cluster and those with fewer than three into another. However, instead of two clusters, the algorithm identified three distinct clusters.

The first cluster consisted of embryos that had three or more cotyledons and a clearly measurable body length. The second cluster included embryos that had fewer than three cotyledons but still exhibited a positive body length. Meanwhile, the third cluster was distinct in that it contained embryos with fewer than three cotyledons but with either negligible or no measurable body length. This unexpected outcome suggests that the algorithm did not solely rely on cotyledon count as a classification criterion. Instead, it recognized body length as a significant feature, leading to an additional cluster. This highlights the algorithm's sensitivity to morphological differences beyond the explicitly defined constraints, demonstrating its ability to detect patterns in the data that might not have been pre-determined by the user. While pairwise constraints in Semi-Supervised Learning (SSL) effectively guide clustering towards desired outcomes, UL maintains the flexibility to identify other differentiating features, ensuring a balance between precision and discovery in clustering results.

2.4 Auto-clustering: Application of Hierarchical clustering algorithm

Unlike the k restriction in K-means or pair constraints in SSL, Hierarchical clustering works by gradually merging the most similar clusters based on a calculated distance between them. Initially, each data point is treated as its own cluster, and the algorithm measures the distance between every pair of points using a chosen metric, such as Euclidean distance, which calculates how far two points are from each other in a multidimensional space. These distances are stored in a matrix, where each value represents the similarity between two points. At each step, the two closest clusters are merged, and the distance between this new cluster and the remaining clusters is updated based on a selected linkage criterion. Single linkage considers the shortest distance between any two points in different clusters, complete linkage takes the farthest distance, and average linkage calculates the mean distance between all points in the two clusters. In image clustering, these distances are based on extracted features such as pixel intensities, shapes, or textures. The merging process continues until

all data points form a single cluster, a concept extensively discussed in [20], where hierarchical clustering was applied to biomedical image analysis for improved pattern recognition. The results are visualized in a dendrogram, a tree-like diagram that represents the sequence of merges, enabling an interpretable clustering structure. By setting an appropriate threshold on this dendrogram, the final number of clusters can be determined dynamically, eliminating the need for predefined cluster counts. This adaptability makes hierarchical clustering particularly useful for complex datasets, such as embryo images, where natural groupings may not be known in advance.

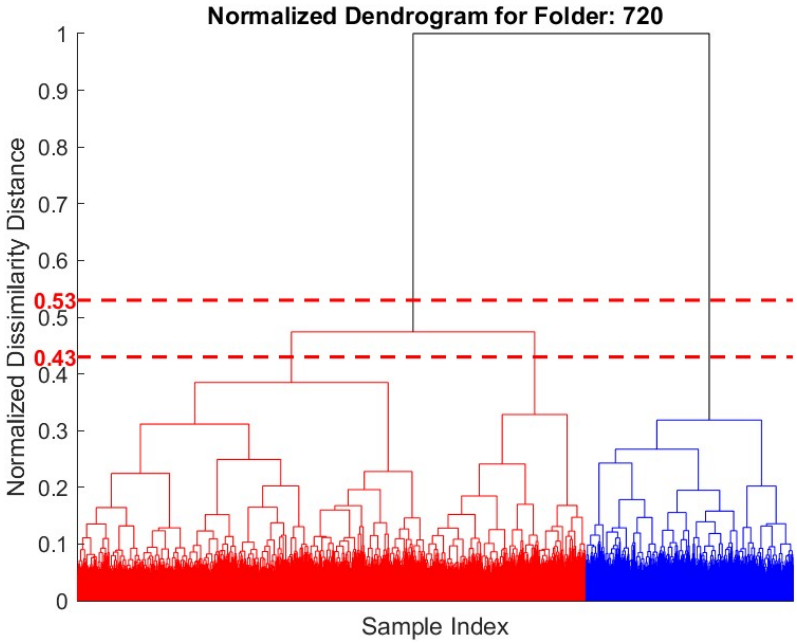


Fig. 6 The Image of a Dendrogram used to visualize the hierarchical clustering process of 720 grayscale embryo images. A threshold of 0.8 was chosen as point of inquiry of cluster share and segregation qualitatively and quantitatively.

For this study, a total of 720 grayscale embryo images were clustered using hierarchical clustering. At a threshold level of 0.8, two distinct output clusters were observed, as illustrated in Fig. 6. The first cluster consisted of 455 images, of which 421 were found out by visual inspection as Accept. In contrast, the remaining 34 were found as Reject. Similarly, the second cluster comprised 265 images, with 206 found out by visual inspection as Reject and 59 found as Accept. These results illustrate that, in contrast to K-Means, hierarchical clustering automatically generated two clusters, which aligned closely with the Accept and Reject categories. The clustering accuracy (compared to human based segregation) was 93% for the first cluster and 78% for the second. This outcome is similar to the results from the K-Means algorithm when k=2, demonstrating that hierarchical clustering can effectively separate the two primary categories without requiring a predefined number of clusters. The approach's

flexibility in not specifying k and its ability to group the images based on inherent data characteristics adaptively make it a useful method in this context.

This algorithm was further employed to cluster and analyses three distinct pre-processed input datasets comprising 500, 3500, and 4089 images, respectively, as illustrated in Fig. 7. For each dataset, dendrograms were constructed to visualize the hierarchical clustering structure. Two types of thresholds were selected for analysis. Lower thresholds were chosen to investigate the merging and de-merging behavior. Specifically, thresholds (normalized dissimilarity distances) of 0.43 and 0.53 were used to examine the cluster states before and after merging, respectively. Higher thresholds, those exceeding 0.75, were employed to assess feature-based segregation qualitatively and cluster share quantitatively. In this analysis, a higher threshold of 0.8 was selected to study the clustering behavior in greater detail.

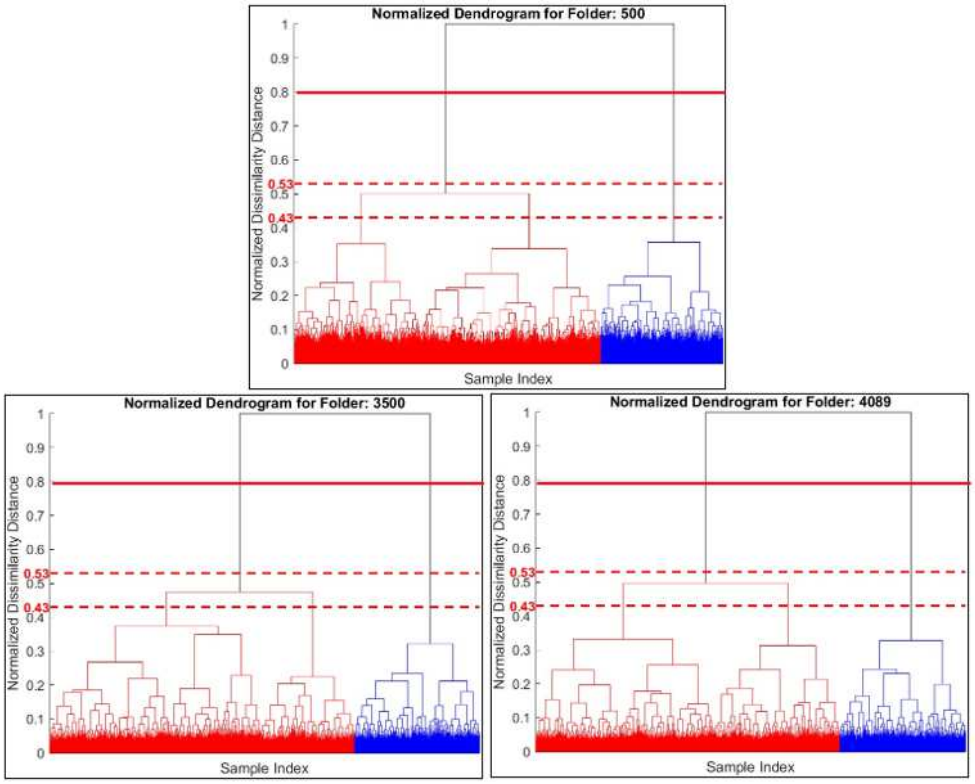


Fig. 7 Sets of Dendrogram plotted for hierarchical clustering done on 500, 3500 and 4089 images of embryos distinctively at bulk. Dendrogram is studied at higher thresholds (0.8) and at lower thresholds (0.53 and 0.43) to study cluster share and merging behavior.

At the higher threshold of 0.8, two clusters were consistently obtained for all input sets, whether comprising 500, 3500, or 4089 images, as depicted in the first row of

Fig. 8. The cluster quality, evaluated based on parameters such as body length, distinct positive cotyledons, and cluster segregation, was found to be consistent across all input datasets. This conclusion was drawn following a detailed analysis that tracked embryo image through their specific file names across thresholds and input datasets. At lower thresholds, visual observations indicated that the primary differentiating factor, embryo body length, led to the formation of three clusters at a threshold of 0.43. The clusters merged at approximately 0.53, as the algorithm increasingly prioritized similarities over distinguishing criteria while progressing from lower to higher threshold levels. These similarities included finite body length and distinct cotyledon characteristics, ultimately leading to a single merged cluster. Notably, the quality and cluster share of the second cluster, designated as Reject, remained consistent at the 0.53 threshold. The second row of Fig. 8 (Lower Thresholds) illustrates three representative images belonging to three clusters for each image data set (500, 3500, 4089). The first two images, despite exhibiting prominent cotyledons and body structure, differ in body length, justifying their classification into two separate clusters. The third image lacks either a prominent body structure or cotyledons and is thus categorized into the third cluster (Reject). At higher thresholds, the image datasets of 500, 3500 and 4089 do not show any noticeable difference in clustering patterns. It is observed that three clusters obtained at a lower threshold are consistently merged to get only two clusters.

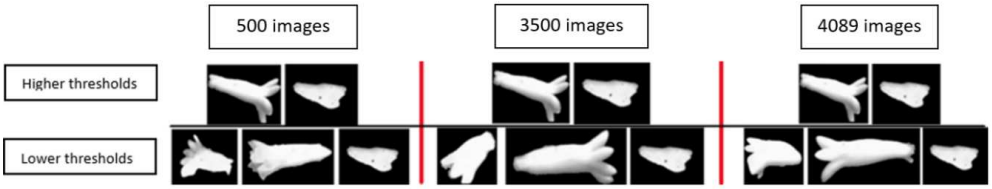


Fig. 8 Each embryo image represents a cluster formed taking input pool of images of 500, 3500 and finally all 4089 images at bulk for lower thresholds before merging and at higher thresholds respectively. Three clusters (shorter accepted embryos, longer accepted, and reject categories).

This consistency suggests that just processing a dataset of 500 images is sufficient to represent the clustering characteristics of the entire dataset. This implies that one may not need a large volume of image data to perform the clustering process successfully.

The graph plotted in Fig. 9 illustrates the variation in the number of clusters formed as a function of images in a given dataset. The number of images in a dataset is plotted on the x-axis (ranging from 100 to 4089 images) and the corresponding number of clusters formed are represented on the y-axis. The plot reveals that, irrespective of the number of iterations (repetitive random extraction of images to group them as folders for further clustering) as represented on x axis of Fig. 9 were performed, two iterations of clustering, depicted by red and blue lines, exhibit convergence after processing 500 images. Beyond this point, both lines remain constant up to 4089 images.

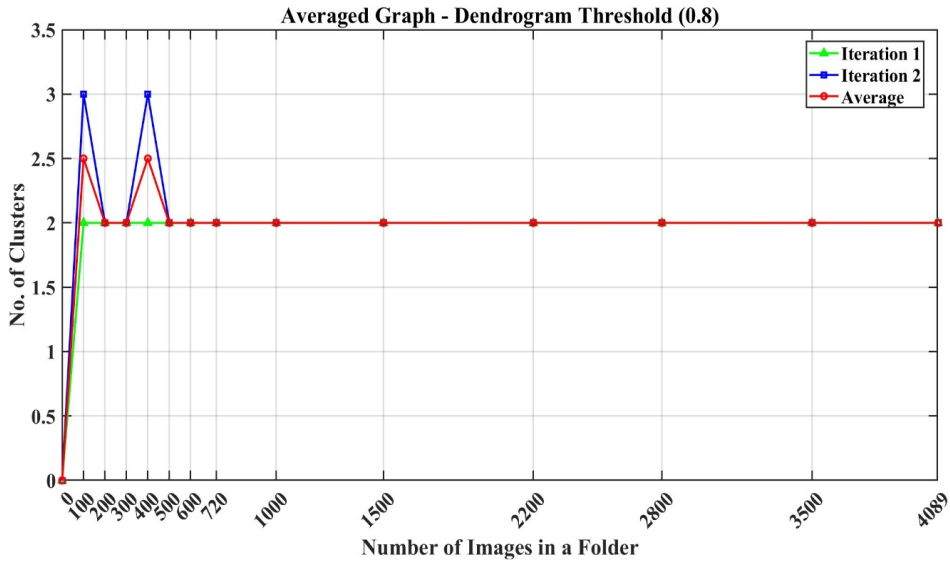


Fig. 9 The graph plotted between no. of clusters procured for no. of images in the dataset's input pool. green and blue points denote different iterations run, whereas the red denotes the averaged-out iteration for better visualization.

This quantitative observation aligns with the qualitative inferences derived from dendrogram analysis, thereby validating the consistency of the clustering behavior across the dataset.

2.5 Online classification of embryo images using combined UL and SL technique

Unlike SL models, most traditional UL and SSL algorithms struggle with real-time clustering of images without incremental updates or online adaptations. A robust system capable of processing a grayscale image as input and classifying it into "acceptable" or "rejectable" categories for subsequent research or implantation is needed. This section details a code architecture, as shown in Fig. 10, that integrates Constrained Unsupervised Learning (CUL) for clustering with the SL SVM technique for classification. CUL initially clusters the image pool based on prominent features without predefined labels. SL then trains on these clusters to classify new test images, ensuring alignment between clustering features and classification objectives. The design of the code architecture is driven by the need to establish a structured clustering framework that mirrors human perception of feature clusters. This allows the algorithm to classify new, unseen images efficiently by mapping them to these learned clusters. Such an approach ensures rapid, real-time categorization while maintaining adaptability for future datasets, making it suitable for diverse classification tasks beyond the current study.

The algorithm is divided into several events.

In Event 1, the clusters generated are identical to those obtained in the preceding SSL/CUL section, as the code is the same in every aspect, including feature extraction,

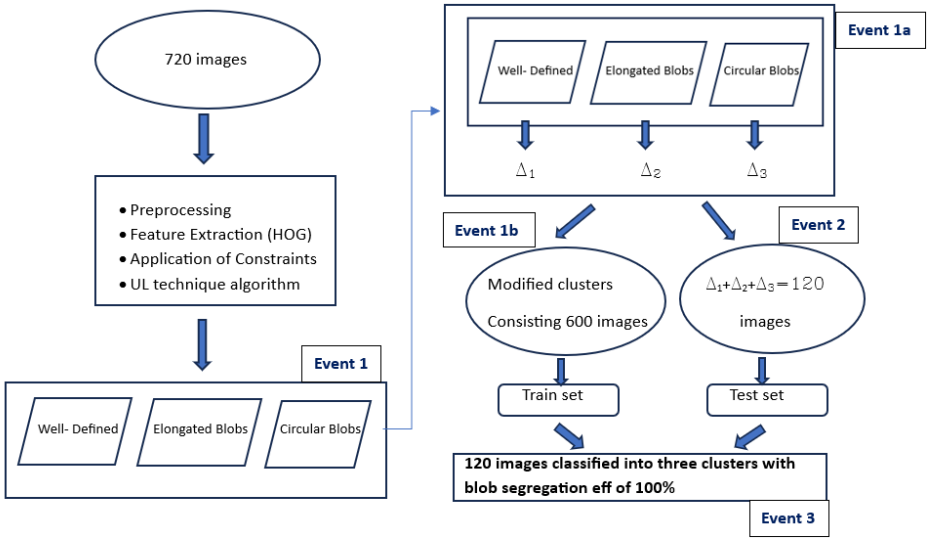


Fig. 10 The algorithm designed to use CUL technique for clustering and SL SVM technique for classification.

dimension reduction, application of constraints, and clustering technique. The three resulting clusters are well-defined, circular blobs and elongated blobs Following Event 1, in Event 1a, 120 images, or ∼16.7%, are randomly extracted from the three clusters as Δ_1, Δ_2, and Δ_3 to serve as test images for Event 2. In Event 1b, the SVM technique is used to train a SL model with the remaining 600 images distributed across the three clusters. For the SL model, the data clustered into the three categories by the UL method serve as the labelled data. In Event 2, the 120 extracted images are then used as a test data set in the SL SVM model. Each image is labelled according to the cluster from which it was taken in Event 1a. Finally, in Event 3, after training the SL model and classifying the 120 images, the classification results are compared with the clustered results. This is easily done by tracking the classified image name string with the origin of the images from the pre-clustered folders. A 100% correct classification of Circular Blobs was achieved with the UL-SL technique. Additionally, classification techniques see these as folders containing failed embryos. The advantage is that the SL classification can effectively handle the on-line decision-making process and provide precise guidance on the next set of actions to be taken on specific embryos.

The trained SL model is applied to the 120 images, and Fig. 11 gives an idea of the non-dimensionalized average processing time taken by an image. Each test image underwent the pipeline stages of read, preprocess, feature extraction, classification and finally, save (label verification and output storage). On an AMD Ryzen 5 5600H Processor with Radeon Graphics, 3301 MHz, 6 Core(s), 12 Logical Processor(s) 8GB SSD RAM system running MATLAB R2022b, the average runtime per test image was 0.00408 seconds. This efficiency suggests even faster classification speeds on higher-performance hardware, making the pipeline suitable for real-time applications. To evaluate the classification performance, a set of test images was processed to compare supervised learning (SL) and morphological classification methods. HOG features were

extracted from each test image, and SL classification times were recorded using a pre-trained SVM classifier. Concurrently, morphological classification times were measured by applying predefined morphological operations, simulating typical processing. The average percentage by which SL was faster than the morphological method was calculated to be 32.3%, highlighting the efficiency of SL in this context. The integration of CUL with SL demonstrates a seamless architecture for addressing human bias during the segregation of embryos and real-time deployment challenges of UL techniques. By clustering images based on dominant morphological features and using these clusters to train SL models, the system provides a robust framework for evaluating embryos in situ, on a large scale.

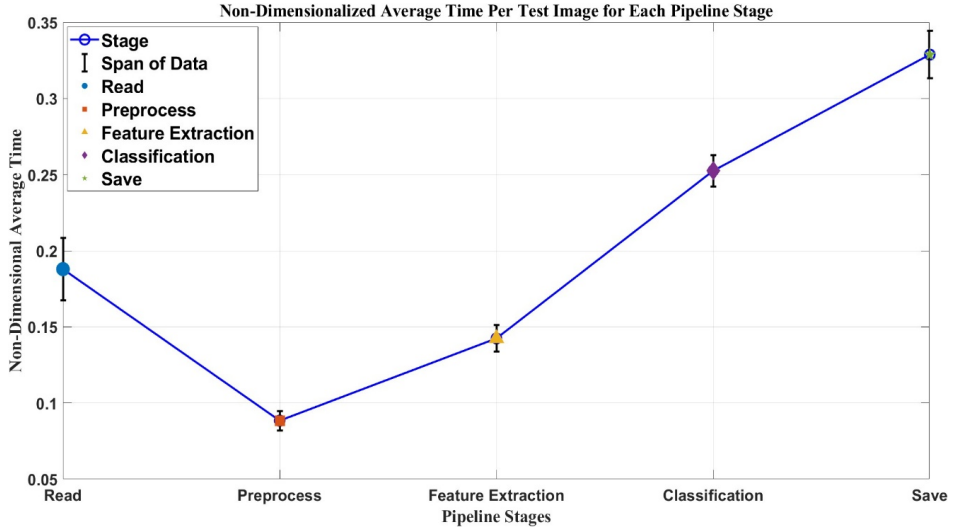


Fig. 11 Non-dimensionalized average time per test image with standard deviation over the data per pipeline stage plotted for the testing process using SVM (SL) technique in the algorithm designed and discussed in Fig. 3.

3 Conclusion & Future Work

Integrating constrained unsupervised learning (CUL) within a supervised learning (SL) framework offers advantages and limitations in clustering tasks. Pairwise constraints enhance targeted feature isolation but restrict the algorithm's capacity for autonomous pattern discovery. In contrast, hierarchical agglomeration and dendrogram-based visualizations provide a flexible, data-driven approach, allowing automated cluster formation without predefined constraints. The balance between guided clustering and autonomous feature discovery, which can be extended to various classification and clustering problems beyond the studied application is essential. The methodology demonstrated here provides a structured way to refine clustering outcomes, reducing reliance on manual labelling while improving classification accuracy when compared to traditional morphology and supervised learning processes. This

approach can be applied to any dataset where feature isolation and efficient clustering are critical, making it valuable for fields that require automated, high-precision clustering. Future work will explore using Hyperspectral Imaging (HSI) instead of conventional RGB imaging to enhance feature extraction and clustering accuracy. By leveraging the rich spectral information provided by HSI, the classification process can be refined to distinguish subtle variations that may not be captured through standard imaging techniques. The accuracy of these algorithm-generated clusters will be evaluated against experimentally tagged groups, providing deeper insights into the reliability and effectiveness of ML-driven classification. Ultimately, this study aims to establish HSI-based machine learning as a scalable, real-time solution for automated classification tasks.

Declarations

Funding: This research received no external funding.

Conflict of interest: The authors declare no competing interests.

Ethical approval: Not applicable.

Consent to participate: Not applicable.

Consent to publish: Not applicable.

Data availability: The datasets used and analyzed during the current study are available from the corresponding author upon reasonable request.

Authors' contributions: All authors contributed to the study conception, methodology, and manuscript preparation.

References

[1] Sevillano, I., Antón-Fernández, C., Søgaard, G., Astrup, R.: Improved forest management for increased carbon sequestration: An assessment of the most prominent approaches in norway. Journal of Environmental Management **375** (2025) https://doi.org/10.1016/j.jenvman.2025.124333

[2] Aspinwall, M.J., McKeand, S.E., King, J.S.: Carbon sequestration from 40 years of planting genetically improved loblolly pine across the southeast united states. Forest Science **58**(5), 446–456 (2012) https://doi.org/10.5849/forsci.11-058

[3] FAO: FAO assessment of forests and carbon stocks, 1990–2015: Reduced overall emissions, but increased degradation (2015). http://unfccc.int/parties_and_observers/

[4] McKendry, P.: Energy production from biomass (part 1): overview of biomass. Bioresource Technology (2002)

[5] Zhao, J., Lange, H., Meissner, H.: Estimating carbon sink strength of norway spruce forests using machine learning. Forests **13**(10) (2022) https://doi.org/10.

3390/f13101721

[6] Rytter, L., *et al.*: Increased forest biomass production in the nordic and baltic countries - a review on current and future opportunities. Silva Fennica (2016) https://doi.org/10.14214/sf.1660

[7] Jyrkiikangas, H., Peltola, H.: Managing Forest Ecosystems: Forest Bioeconomy and Climate Change, (2017)

[8] Aidun, C.K., Egertsdotter, U.: Se fluidics system. In: Somatic Embryogenesis, pp. 211–227 (2018). https://doi.org/10.1007/978-3-319-89483-6_16

[9] Le, K.C., Weerasekara, A.B., Ranade, S.S., Egertsdotter, E.M.U.: Evaluation of parameters to characterise germination-competent mature somatic embryos of norway spruce (picea abies). Biosystems Engineering **203**, 55–59 (2021) https://doi.org/10.1016/j.biosystemseng.2020.12.013

[10] Jamuna, K.S., Karpagavalli, S., Vijaya, M.S., Revathi, P., Gokilavani, S., Madhiya, E.: Classification of seed cotton yield based on the growth stages of cotton crop using machine learning techniques. In: ACE 2010 - International Conference on Advances in Computer Engineering, pp. 312–315 (2010). https://doi.org/10.1109/ACE.2010.71

[11] Vlasov, A.V., Fadeev, A.S.: A machine learning approach for grain crop's seed classification in purifying separation. In: Journal of Physics: Conference Series (2017). https://doi.org/10.1088/1742-6596/803/1/012177

[12] Niazian, M., Sadat-Noori, S.A., Abdipour, M., Tohidfar, M., Mortazavian, S.M.M.: Image processing and artificial neural network-based models to measure and predict physical properties of embryogenic callus and number of somatic embryos in ajowan. In Vitro Cellular and Developmental Biology - Plant **54**(1), 54–68 (2018) https://doi.org/10.1007/s11627-017-9877-7

[13] Arab, M.M., Yadollahi, A., Shojaeiyan, A., Ahmadi, H.: Artificial neural network genetic algorithm as powerful tool to predict and optimize in vitro proliferation mineral medium for g x n15 rootstock. Frontiers in Plant Science **7** (2016) https://doi.org/10.3389/fpls.2016.01526

[14] Jamshidi, S., Yadollahi, A., Ahmadi, H., Arab, M.M., Eftekhari, M.: Predicting in vitro culture medium macro-nutrients composition for pear rootstocks using regression analysis and neural network models. Frontiers in Plant Science **7** (2016) https://doi.org/10.3389/fpls.2016.00274

[15] Egertsdotter, U., Ahmad, I., Clapham, D.: Automation and scale up of somatic embryogenesis for commercial plant production, with emphasis on conifers. Frontiers in Plant Science (2019) https://doi.org/10.3389/fpls.2019.00109

[16] Chatterjee, P., Weerasekara, A., Egertsdotter, U., Aidun, C.: Machine learning based sorting of somatic embryos for in-line processing in automated se fluidics system. Agricultural Research and Technology: Open Access Journal **28**(1) (2024) https://doi.org/10.19080/ARTOAJ.2024.28.556399

[17] Sangeetha, D., Deepa, P.: Efficient scale invariant human detection using histogram of oriented gradients for iot services. In: Proceedings - 2017 30th International Conference on VLSI Design and 2017 16th International Conference on Embedded Systems, pp. 61–66 (2017). https://doi.org/10.1109/VLSID.2017.60

[18] Lu, Z., Leen, T.K.: Semi-supervised Clustering with Pairwise Constraints: A Discriminative Approach

[19] Masud, M.A., Huang, J.Z., Zhong, M., Fu, X.: Generate pairwise constraints from unlabeled data for semi-supervised clustering. Data and Knowledge Engineering **123** (2019) https://doi.org/10.1016/j.datak.2019.101715

[20] Eric, O.U., Michael, O.O.: Overview of agglomerative hierarchical clustering methods. British Journal of Computer, Networking and Information Technology **7**(2), 14–23 (2024) https://doi.org/10.52589/bjcnit-cv9poogw