Comparative Analysis of Machine Learning Techniques in the Classification of Pili *(Canarium ovatum Engl.)* Fruit Varieties

Leo Constantine S. Bello
Ateneo de Naga University
Naga City, Camarines Sur
lbello@gbox.adnu.edu.ph

Joshua C. Martinez
Ateneo de Naga University
Naga City, Camarines Sur
joshuamartinez@gbox.adnu.edu.ph

ABSTRACT

Pili is one of many Philippine fruit trees that are endemic in the country and are predominantly found in the Bicol Region. Existing means of classifying Pili are available and highly accurate but are known to be invasive, destructive, costly, and require a highly technical background. Non-invasive means through Computer Vision techniques have been applied to some fruits, but no study has applied it yet in classifying pili. This study aimed to provide a comparative analysis of four (4) machine learning algorithms: LDA, k-NN, SVM, and CNN, commonly used in image classification, and assess which is best suited for further study and application.

KEYWORDS

Pili, Canarium ovatum Engl., Comparative Analysis, Image Processing, Variety Classification

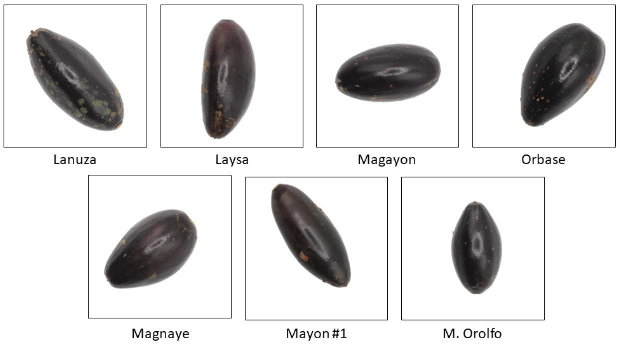


Figure 1: Visual feature differences among pili fruit varieties

1 INTRODUCTION

The Philippines is home to abundant fruit and nut trees. It is said to be the center of diversity for several fruit trees that bear edible nuts. *Canarium ovatum Engl*, also known as Pili, is one of the most important edible nut-bearing trees. This species of the Canarium genus is endemic in the country, and its high density of population distribution and growth is restricted to areas relatively close to its center of origin, the Bicol Region. Its fruit kernel distinguishes Pili. Pili has been designated as a top priority, high-value, and high-impact fruit crop due to the government's research and development efforts, joining the ranks of mango, durian, lanzones, rambutan, banana, papaya, and citrus.

The Pili tree is known as the "Tree of Hope" because of its many uses, from sap to roots. The most important part of the fruit, the kernel, is processed into pastries and confectioneries, and the whole kernel is now emerging in the local and international markets. Its pulp is considered a delicacy in some areas outside the region, and its shells can be used to make charcoal fuel or handicrafts. Its nut oil has potential applications in the food and pharmaceutical industries and is in high demand for domestic and international exports.

Pili has been identified as a highly variable species. The trees differed in growth habits, fruiting season, yield, response to asexual propagation, stem diameter, leaf size, number of flower clusters/shoot, and flowering period. One of the species' most noticeable features is its fruits, which vary in shape, color, weight, thickness (pulp and shell), flavor, kernel size, and content. The fruit is 4-6 cm in size and comes in elliptical, oblong, oval, and obovate shapes. Its pulp turns from green to dark purple to nearly black as it ripens. Saturated and unsaturated fatty acid percentages in the oil vary, as do the percentages of filled nuts and kernel recovery in seedling trees. Calcium, potassium, zinc, and phytic acid concentrations vary from variety to variety [6]. When identifying fruit varieties, someone with sufficient knowledge and expertise could determine and distinguish variations based on their visual characteristics, as shown in Fig.1.

Fruit variety classification is critical for producers and farmers to ensure the purity and crop output of the variety. Furthermore, because of their potential to increase productivity and profitability, saplings and fruits from the recommended variety for production were deemed more expensive. However, the morphological and color characteristics of pili fruit from various types are strikingly similar, with significant overlap. A person with little to no knowledge of or experience with different types may have difficulty identifying one. These circumstances may present an ideal opportunity for any enterprising individual to offer low-quality pili variants illegally, as high-quality increases profit margins. Furthermore, some orchards grow a diverse range of pili varieties to adapt to changing climates, harvest seasons, and other agronomic factors. Many small orchards grow diverse fruits with varying demand and market value, which can complicate local post-harvesting due to unintentional and fraudulent mixing of different fruit varieties. This problem is not new for other crop groups such as rice [8] and corn [4].

Historically, Pili variations were identified solely by visual inspection. Botanists use plant traits as identification keys, then conduct sequential and adaptive research to identify varied plant kinds. The technique focuses on answering questions about pili fruit qualities, including form, color, length, and width. Consistently focusing on unique traits refines the species pool. Accurately responding to inquiries leads to the desired diversity. Additionally, farmers and producers rely on manual inspection and sorting due to cost-effectiveness and limited professional availability. The classification outcome may be affected by investigator competence and subjectivity, which can lead to discrepancies, workload, and tiredness. Identifying traditional varieties is difficult for the general public, but considerably more difficult for botanical experts such as conservationists, farmers, foresters, and product designers who face frequent problems. Identifying a certain type might be challenging even for professionals. Drawbacks of this method include high error rates, low precision, and significant processing time, particularly for specific types.

High-performance analytical procedures that are commonly used include liquid chromatography, gas chromatography-mass spectrometry [9], seed protein electrophoresis [11], and DNA molecular markers [12]. Despite their high accuracy, most of these technologies are invasive, harmful, hazardous to human health, time-consuming, sophisticated, and expensive, with little chance of being repeatable in the future. As a result, it is critical to use a safe, non-destructive, and accurate automated system for variety classification. These non-destructive automation processes can save money by increasing efficiency while reducing subjectiveness caused by human experts. Numerous studies in the field of agronomy have demonstrated the use of non-destructive plant species classification techniques such as magnetic resonance imaging [16], electronic tongue [10], acoustic method [15], electronic nose [1], and computer vision. Computer vision and image processing are two methods for classifying crops that are both low-cost and provide significant analytical and computational power. Image pre-processing, segmentation, feature extraction, and classification are the steps in computer vision-based classification, with feature extraction significantly impacting both classification accuracy and classification precision.

Several studies already implemented the use of computer vision techniques in the classification of crops and plants. In this study, the researchers would like to explore some notable machine learning techniques such as Linear Discriminant Analysis (LDA), k-Nearest Neighbor (k-NN), Support Vector Machine (SVM), and Convolutional Neural Networks (CNN) wherein they performed well in classifying fruit subjects [2–5, 7, 13, 14], and assess its viability in accurately classifying the pili fruit according to its variety.

2 METHODOLOGY

2.1 Data Gathering

In this study, a total of 1,400 pieces of ripe pili fruit samples from seven (7) varieties were harvested at the Albay Research and Development Center, Department of Agriculture Field Office V in Buang, Tabaco City. The sample collection comprised two hundred (200) fruits from each variety, namely, Lanuza, Laysa, Magayon, Magnaye, Mayon #1, M. Orolfo and Orbase. All fruits collected were from the varieties labeled parent trees with the assistance of a resident agriculturist. The harvest was done one tree at a time. It was to ensure no occurrences of mixing up of fruit samples.The collection was done in August 2023 during one of the fruit's maturing months.

To augment the existing dataset, the researcher gathered another set of two hundred (200) fruit samples from pili cultivars from four (4) different trees. These species of pili are a product of selective propagation and were not considered as varieties. Selection of the cultivar trees were assisted by the resident agriculturist. These samples are were comprised to form another group of data called 'Cultivars'. This was done to ensure that the model would recognize distinctions and discrimination between varieties and cultivars.

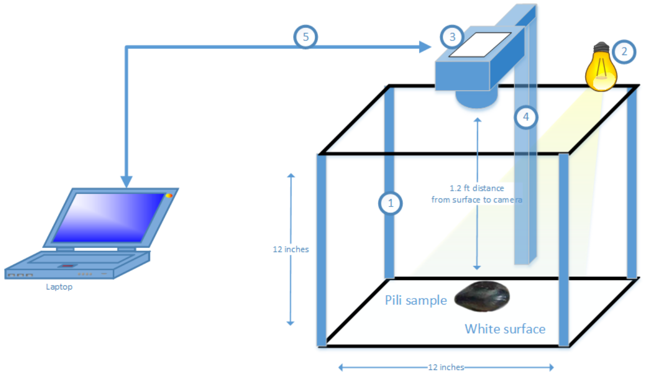


Figure 2: Image Acquisition Setup

2.2 Image Acquisition: Materials and Methods

To capture the images of the fruits, an image acquisition system (Fig. 2) was assembled. The aim was not to implement a final version of the system but a prototype mimicking conditions or environment where samples were to be captured. The system was conceived to acquire images with high contrast between the subject and the background with the aim of capturing image with minimal shadow cast. This system underwent multiple experimentation on the configuration of the camera as well as the lighting before capturing pili samples for further image analysis.

2.3 Image Pre-processing

Captured original images underwent background subtraction to remove unnecessary objects. Afterwards the images were cropped to have equal height and width of 2912 x 2912 pixels centered to the subject. To lessen the storage and memory space during the image analysis,image-set underwent to a process of resizing images to 700 x 700 pixels. This is done through batch processing employing python 3.11.0 and OpenCV's image processing capabilities.

Images were then subjected to background cleaning to remove unnecessary objects such as dirt and foreign objects found in the background. This is followed by replacement to plain white background to provide ease in further image processing activities such as thresholding and binarization which were performed during feature extraction.

To augment the dataset, images were subjected to horizontal flip followed by vertical flip. Each of the flip were stored to a separate folder named according to their respective varieties. To further augment the dataset, images were rotated 90 degrees, 180 degrees and 270 degrees. Considering the original and the augmented images resulted to total of 57,600 images, having 7200 for each variety. Due to computational cost for having 700 x 700-pixel images in feature extraction and other process in image analysis, resizing images to 256x256 pixels were considered in this study.

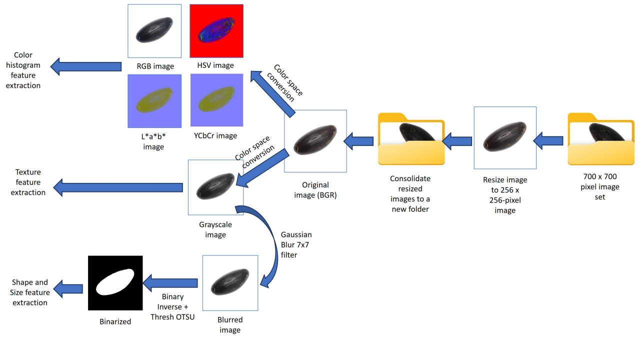

Color histogram feature extraction
RGB image
HSV image
L*a*b* image
YCbCr image
Color space conversion
Color space conversion
Original image (BGR)
Consolidate resized images to a new folder
Resize image to 256 x 256-pixel image
700 x 700 pixel image set
Texture feature extraction
Grayscale image
Gaussian Blur 7x7 filter
Shape and Size feature extraction
Binarized
Binary Inverse + Thresh OTSU
Blurred image

Figure 3: Image pre-processing activities prior to feature extraction.

2.4 Feature Extraction

The next step in developing a classifier model after the image pre-processing is feature extraction. Main and important visual external features for a fruit in general are its color, texture, shape and size. Fig. 3 illustrates further the image pre-processing activities before feature extraction. Extracted features were saved in Comma Separated Values (.csv) files to be loaded during the classifier modelling. It should be noted that features under color, texture, shape and size were hand-crafted and is intended for modelling traditional machine learning algorithms such as, in this study, LDA, SVM and K-NN. Higher level feature extraction from a transfer learning technique was used for the CNN model.

2.4.1 Color feature extraction. Color histogram analysis was performed on different color spaces such as RGB, HSV, L*a*b*, and YCbCr. This is one of the most important tasks in image processing and computer vision tasks. Each of the mentioned color spaces provide valuable insights on the development of the pili fruit. In this study, the features were extracted accordingly and were summarized using Histogram Mean and Standard Deviation.

2.4.2 Texture feature extraction. Three (3) methods were used to extract texture features: GLCM (Grey-level Co-occurrence Matrix), GLRLM (Grey-level Run Length Matrix) and the DWT (Discrete Wavelet Transform). The GLCM extracts the features Contrast, Energy, Homogeneity and Correlation. GLRLM extract the features Short Run Emphasis (SRE), Long Run Emphasis (LRE), and Gray-level Non-Uniformity (GLN). DWT was used to extract Approximate and Detailed Energy, and Shannon Entropy. features

2.4.3 Shape feature extraction. Shape features were extracted after defining the image contours. Shape feature extracted were: Area, Perimeter, Compactness, Roundness, Aspect Ratio, Eccentricity, Solidity, and Hu Moments.

2.4.4 Size feature extraction. To extract the size of the fruit samples, the researcher used the features bounding box dimension (width and height), and the diameter.

2.4.5 High-level feature extraction. To extract the High-level visual features, the weights of ImageNet[1] coupled with VGG168[2] CNN Model was used. Using this pre-trained transfer learning technique, the researcher aimed to extract features that cannot be extracted by hand-crafted way.

2.5 Classifier Modelling

In modeling a non-invasive image-based classifier for pili varieties, three (3) traditional Machine Learning models – Linear Discriminant Analysis (LDA), Support Vector Machine (SVM) and k-Nearest Neighbor (k-NN), and a Deep Learning model – Convolutional Neural Network (CNN) was considered.

2.5.1 Traditional Machine Learning. In loading the dataset in the model, the traditional machine learning used hand-crafted features: Color, Texture, Shape and Size feature vectors in .csv format. These features were loaded individually, and in combination of all four features to test the effectivity of the designed traditional machine learning model.

The feature sets were split into the ratio of 70:15:15, wherein 70 percent of the dataset is the training, 15 percent for the testing, and another 15 percent for the validation. Selection of the data samples were done in random. Due to values inside the datasets were diverse in range, the training, testing, and validation sets were subjected to normalization before loading them into the model. To reduce the dimensionality of the datset, Principal Component Analysis (PCA) was used. In this study, the number of components to be retained was 10. This was to set to balance the trade-off between dimesionality and accuracy (overfitting or underfitting) of the models.

(1) Linear Discriminant Analysis (LDA) - The LDA was set with 'lsqr' solver or Least Squares solution. This was considered due to the 'lsqr' solver can be applied to high-dimesional but sparse data because it leverages sparsity to speed up computation. The shrinkage or the regularization of the model was set to 'auto' where it used the Ledoit-Wolf Lemma. This helped the model in improving its robustness and stability especially with the study's scenario of high dimensional data. Lastly, the number of components were into 'None' this is to preserve the original values of feature values.

(2) Support Vector Machine (SVM) – The models' kernel is set to 'poly' or Polynomial. This is due to the data not linearly separable. The C which is the Regularization parameter is set default value of 1. This is because larger values tend to overfit and lesser values of C tends to underfit the resulting model. The gamma was set to 'scale'. The decision function shape was set to 'ovr' or one-vs-rest strategy of SVM to accommodate multi-class classification. 'ovr' create a total comparison of N X number_of_binary_classifiers, where N is the number of classes. In comparison with 'ovo' (one-vs-one), this strategy has faster training time and works well with imbalances in classes.

[1] Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009.

[2] Tammina, Srikanth. "Transfer learning using vgg-16 with deep convolutional neural network for classifying images." International Journal of Scientific and Research Publications (IJSRP) 9.10 (2019): 143-150.

(3) k-Nearest Neighbor (k-NN) – The k-NN model set the number of neighbors to 30. This is to ensure that the result of the model is a balanced due to lesser value tend to overfit the model. The value of the power parameter p is set to 2 so that the Minkowski distance would be transformed to Euclidean. The algorithm was set to 'auto' to let the classifier decide the most appropriate algorithms from BallTree, KDTree and Brute Force depending on the values passed to the train function. Lastly, the weights were set to 'uniform'. This sets all the neighbors to be weighted equally.

2.5.2 Convolutional Neural Network. This work utilized the embedded data generator of keras to create a dataset from actual images, instead than depending on feature vectors. The original dataset photos were separated into 70:15:15 ratios: 70% for training, 15% for test, and 15% for validation. This yielded 420 original photos per class folder for training, 90 for testing, and 90 for evaluation. This value is multiplied by eight (8) recognized classes. The CNN model was defined with 8 classes, 32 batches, and 100 epochs. Augmentations included rotation, width, height, shear, zoom, horizontal flip, vertical flip, and fill nearest. This increases the randomness and variation of the original dataset.

The CNN base model combined transfer learning from VGG16 with ImageNet weights. Input images were resized to 224 x 224 pixels for VGG16 design. Using sequential design, the CNN model can be developed from input through output, with layers stacked early. The CNN model was augmented with a pre-trained base model. Flattened, this becomes a 1-dimensional vector. A Dense or Fully linked layer with 512 neurons with 'ReLU' activation was added to the original model after flattening. A Dropout of 0.5 was introduced to the model to prevent overfitting. To penalize the loss function, L2 regularization was added to the dense layer with a value of 0.01. In the last Dense layer, eight neurons reflect the number of classes in the classification challenge. Softmax, often used for multi-class classification, was utilized in this layer. The output layer showed anticipated class probabilities for each class.

2.6 Performance Evaluation

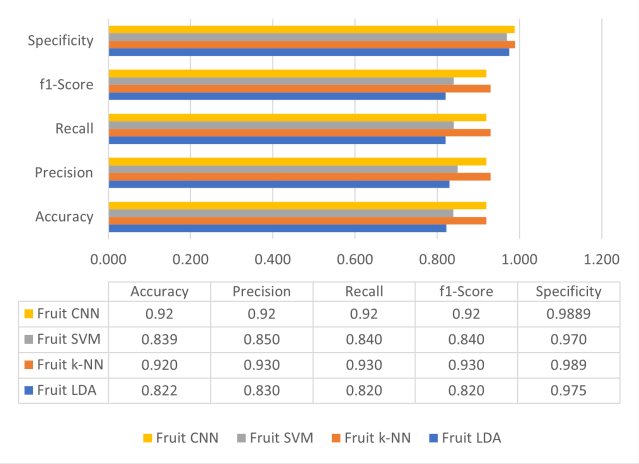
Models produced from LDA, k-NN, SVM, and CNN will be evaluated in terms of Accuracy, Precision, Recall, Specificity, and F-measure. To determine the classifier's performance, the study used k-fold cross validation coupled with kappa statistics. The results of k-fold cross validation will be subjected to confusion matrix to provide detailed picture of the performance of the classifiers.

3 PRELIMINARY RESULTS

3.1 Model-based performance

After the models were trained, they were subjected to an evaluation of their performance. The evaluation involved using the testing and validation portions of the dataset in order to determine their rate of recognition and discrimination of unknown objects. The models were evaluated based on their accuracy, precision, recall, specificity, and F1-score. To be able to determine how well the model performed, k-fold cross validation and kappa statistics performed.

Fig. 4 shows the performance of the models LDA, SVM, and k-NN using the criteria of accuracy, precision, recall, F1-score, and specificity. The highest score for accuracy was gained by CNN with a score of 0.92, and LDA gained the least accuracy score of 0.822. For precision of the models, k-NN gained the highest score of 0.930, and LDA got the least score of 0.830. The same ranking and score were gained for recall and F1-Score. For specificity, the k-NN model gained the highest score of 0.989, and the least was gained by SVM with a score of 0.970.



	Accuracy	Precision	Recall	f1-Score	Specificity
Fruit CNN	0.92	0.92	0.92	0.92	0.9889
Fruit SVM	0.839	0.850	0.840	0.840	0.970
Fruit k-NN	0.920	0.930	0.930	0.930	0.989
Fruit LDA	0.822	0.830	0.820	0.820	0.975

Figure 4: Performance Evaluation results for LDA, k-NN, SVM and CNN in terms of Accuracy, Precision, Recall, f1-Score, and Specificity

4 FURTHER WORKS

The exhaustive analysis of the classifiers conducted in this study will continue to focus on class-based performance, in which the varieties will be evaluated using the same metrics. This is done to understand the dynamics of each fruit's visual features and the factors that influence the analysis's outcome. The classifiers will also undergo k-fold cross-validation and kappa statistics to determine their overall performance. We will implement application software to test the viability and usability of the best-performing classifier on real-world pili fruit samples.

REFERENCES

[1] Takahiro Arakawa, Kenta Iitani, Koji Toma, and Kohji Mitsubayashi. 2021. Biosensors: Gas Sensors. (2021).

[2] Dhiya Mahdi Asriny, Septia Rani, and Ahmad Fathan Hidayatullah. 2020. Orange Fruit Images Classification using Convolutional Neural Networks. In *IOP Conference Series: Materials Science and Engineering*, Vol. 803. IOP Publishing, 012020.

[3] Sumaira Ghazal, Waqar S Qureshi, Umar S Khan, Javaid Iqbal, Nasir Rashid, and Mohsin I Tiwana. 2021. Analysis of visual features and classifiers for Fruit classification problem. *Computers and Electronics in Agriculture* 187 (2021), 106267.

[4] Shima Javanmardi, Seyed-Hassan Miraei Ashtiani, Fons J Verbeek, and Alex Martynenko. 2021. Computer-vision classification of corn seed varieties using deep convolutional neural network. *Journal of Stored Products Research* 92 (2021), 101800.

[5] Lazhar Khriji, Ahmed Chiheb Ammari, and Medhat Awadalla. 2020. Artificial Intelligent Techniques for Palm Date Varieties Classification. *International Journal of Advanced Computer Science and Applications* 11, 9 (2020), 489–495.

[6] Cristopher G Millena, Bernardo A Altavano, and Rosario S Sagum. 2021. Dietary Fiber and Fermentability Characteristics of Different Pili (Canarium ovatum, Engl.) Varieties in the Philippines. *Philippine Journal of Science* 150, 4 (2021), 845–855.

[7] Juan M Ponce, Arturo Aquino, and José M Andújar. 2019. Olive-fruit variety classification by means of image processing and convolutional neural networks. *IEEE Access* 7 (2019), 147629–147641.

[8] Salman Qadri, Syed Furqan Qadri, Abdul Razzaq, Muzammil Ul Rehman, Nazir Ahmad, Syed Ali Nawaz, Najia Saher, Nadeem Akhtar, and Dost Muhammad Khan. 2021. Classification of canola seed varieties based on multi-feature analysis using computer vision approach. *International Journal of Food Properties* 24, 1 (2021), 493–504.

[9] Zhengjun Qiu, Jian Chen, Yiying Zhao, Susu Zhu, Yong He, and Chu Zhang. 2018. Variety identification of single rice seed using hyperspectral imaging combined with convolutional neural network. *Applied Sciences* 8, 2 (2018), 212.

[10] María L Rodríguez-Méndez, C Apetrei, and José A De Saja. 2010. Electronic tongues purposely designed for the organoleptic characterization of olive oils. In *Olives and olive oil in health and disease prevention*. Elsevier, 525–532.

[11] Simona Rogl and Branka Javornik. 1996. Seed protein variation for identification of common buckwheat (Fagopyrum esculentum Moench) cultivars. *Euphytica* 87, 2 (1996), 111–117.

[12] Carlo Miguel C Sandoval, Evelyn Mae Tecson-Mendoza, and Roberta N Garcia. 2017. Genetic diversity analysis and DNA fingerprinting of pili (Canarium ovatum Engl.) using microsatellite markers. *Philippine Agricultural Scientist* 100, 1 (2017).

[13] Sean Huey Tan, Chee Kiang Lam, Kamarulzaman Kamarudin, Abdul Halim Ismail, Norasmadi Abdul Rahim, Muhamad Safwan Muhamad Azmi, Wan Mohd Nooriman Wan Yahya, Goh Kheng Sneah, Moey Lip Seng, Teoh Phaik Hai, et al. 2021. Vision-Based Edge Detection System for Fruit Recognition. In *Journal of Physics: Conference Series*, Vol. 2107. IOP Publishing, 012066.

[14] Alper Taner, Yeşim Benal Öztekin, and Hüseyin Duran. 2021. Performance analysis of deep learning CNN models for variety classification in hazelnut. *Sustainability* 13, 12 (2021), 6527.

[15] Yossi Yovel, Matthias Otto Franz, Peter Stilz, and Hans-Ulrich Schnitzler. 2008. Plant classification from bat-like echolocation signals. *PLoS Computational Biology* 4, 3 (2008), e1000032.

[16] Yifan Zhou, Raphaël Maître, Mélanie Hupel, Gwenn Trotoux, Damien Penguilly, François Mariette, Lydia Bousset, Anne-Marie Chèvre, and Nicolas Parisey. 2021. An automatic non-invasive classification for plant phenotyping by MRI images: An application for quality control on cauliflower at primary meristem stage. *Computers and Electronics in Agriculture* 187 (2021), 106303.