IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 April 2025, accepted 5 May 2025, date of publication 12 May 2025, date of current version 21 May 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3569213

RESEARCH ARTICLE

An Efficient Encoding Spectral Information in Hyperspectral Images for Transfer Learning of Mask R-CNN for Instance Segmentation of Tomato Sepals

ŽELJANA GRBOVIĆ[1], MARKO PANIĆ[1], VLADAN FILIPOVIĆ[1], SANJA BRDAR[1], HENDRIK DE VILLIERS[2], MANON MENSINK[2], AND ANEESH CHAUHAN[2]

[1]BioSense Institute, University of Novi Sad, 21000 Novi Sad, Serbia
[2]Wageningen Food and Biobased Research, Wageningen University and Research, 6708 PB Wageningen, The Netherlands

Corresponding author: Željana Grbović (zeljanagrbovic@biosense.rs)

This research was funded by the Centre of Excellence for Advanced Technologies in Sustainable Agriculture and Food Security (ANTARES), from the Horizon 2020 research and innovation program under the European Union's Horizon 2020 research and innovation program (SGA-CSA. No. 739570 under FPA No. 664387 and supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia (Grant No. 451-03-136/2025-03/200358).

ABSTRACT The most vulnerable parts of tomatoes are the tips of the sepals, which are the primary entry points for fungal spores. Their precise segmentation within hyperspectral images (HSIs) plays a pivotal role in the development of automated and non-destructive systems for assessing tomatoes' sensitivity to fungal infections. This research addresses the critical need for encoding spectral information in hyperspectral imaging to enhance the efficiency of such automated systems. We investigate four different techniques: Principal Component Analysis (PCA), Independent Component Analysis (ICA), Probabilistic Principal Component Analysis (PPCA), and Non-Negative Matrix Factorization (NMF), to perform transfer learning for tomato sepal instance segmentation using models previously trained on RGB images. A comparative analysis of three Mask Region-based Convolutional Neural Network (Mask R-CNN) backbone models is conducted: the Faster R-CNN, Deformable ConvNet, and Feature Pyramid Network (FPN) on spectral-encoded HSIs of the Brioso tomato variety. The Mask R-CNN with FPN, integrated with the NMF technique achieved the highest level of accuracy, yielding a Mean Average Precision (mAP) of 94.05%. Furthermore, on the second dataset, which included an additional three tomato varieties: Capricia, Provine, and Sao Paolo, the same model achieved mAP score of 86.42% across all tomato varieties with only a single false positive detection. Additionally, we incorporated a custom convolutional layer initialized it with estimated NMF coefficients, and achieved a mAP score of 87.40%. This demonstrates the potential of integrating spectral information encoding with trained deep learning-based instance segmentation models to enable robust and accurate transfer learning for automated agricultural food quality assessments.

INDEX TERMS Hyperspectral imaging (HSI), encoding spectral information in HSI, deep learning, transfer learning, instance segmentation, tomato.

I. INTRODUCTION

Tomatoes (*Solanum lycopersicum*) are a key commercial commodity with a high susceptibility to post-harvest fungal infections. This susceptibility plays an influential role in investments to improve tomatoes' growth, storage, and supply chain conditions. Many factors contribute to the level of a tomato's quality and its vulnerability to pathogenic fungi. These factors include the age of the plant, the number of harvests, humidity, temperature, and fruit handling during post-harvest logistics such as transport, packing, and storage conditions [1], [2]. The most fragile parts of a tomato are the sepal tips, which are usually the entry point for fungal spores. Additionally, pathogenic fungal infections can be spread to

The associate editor coordinating the review of this manuscript and approving it for publication was Yongqiang Zhao.

other parts of the tomato, such as the stem, calyx, or tomato skin [3]. The most common pathogenic fungi that infect tomatoes are Penicillium, Aspergillus, and Mucor [4], [5], [6]. Early prediction of the level of tomato susceptibility to infection by pathogenic fungi would allow timely measures to prevent post-harvest losses, which sometimes reach up to 30% of the yield [7], adjust storage and packaging conditions, and optimize decision-making that extends the shelf life of tomatoes. Frequent and time-consuming visual inspections can help prevent the presence of infected tomatoes on retail shelves. Since these inspections are not precise enough, especially in the earliest stages of infection [8], [9], [10], the development of non-destructive methods for estimating the tomato's susceptibility before the consequences of a pathogenic fungus infection become visible is crucial for the post-harvest supply chain. This research has significant practical implications for food quality assessments. By improving the detection of infected tomatoes, our findings can enhance the accuracy and efficiency of quality control processes, ensuring that only the best produce reaches retail shelves. While aforementioned studies have explored various methods for detecting tomato diseases, significant gaps remain in the ability to efficiently identify infected produce at the point of sale. Additionally, most previously stated research has focused on general disease detection without considering the specific implications for food quality assessments in precision agriculture. Precision farming, leveraging technologies like IoT and machine learning, enables farmers to optimize crop yields and reduce costs by providing data-driven insights on resource requirements [11], [12], [13], [14]. Building on the advancements in precision farming, this study addresses these gaps by introducing a more efficient, automated approach based on hyperspectral imaging (HSI) for a precise detection of sepals that is a crucial step for identifying infected tomatoes in retail settings, with potential applications for real-time quality control. The HSI has emerged as a highly effective, non-invasive technique that surpassed traditional RGB images in acquiring more detailed visual information by having tens of contiguous spectral bands in the non-visible part of the electromagnetic spectrum of light [15]. With machine learning (ML) methods, both unsupervised and supervised, the HSI is used to develop more accurate and precise models for plant disease detection. In [16] the authors found that the near-infrared (NIR) spectral range (700 nm−1300 nm) provided more valuable information for tomato disease detection compared to visible color images (350 nm−700 nm). They employed an extreme learning machine (ELM) classifier model for the detection and classification of fungal diseases on tomato leaves. The importance of the NIR spectral range is also supported by [17] where authors developed an approach that included minimum noise fraction (MNF) transformation [18], multi-dimensional visualization [19], pure pixels endmember selection [20] and spectral angle mapping (SAM) [21] to process the hyperspectral images for identification of diseased tomato plants. The sensitivity of hyperspectral imaging to specific forms of cell damage has also been highlighted in various studies. For instance, researchers have demonstrated its efficacy in detecting the bruise regions on kiwi [22] that were extracted from the images generated by the principal component analysis (PCA) using parallelepiped classification [23]. For the detection of bruises on blueberries, the authors from [24] used support vector machine (SVM) classification [25]. In [26], HSI data were analyzed using partial least squares-discriminant analysis (PLS-DA) and SVMs with data fusion principles. They merge three-level strategies at the data, feature, and decision levels for blueberry bruising detection. In [27], two classifiers of spectral angle mapping (SAM), multinomial logistic regression (MLR), and classification decision trees were used to verify the results of the identification of blueberry fruit on plants. To discriminate early disease in blueberries, [28] employed PLS-DA models on two types of spectral range (i.e., the full wavelength range of 400 nm − 1000 nm and the effective spectral range of 685 nm − 1000 nm), showing that the effective spectral range provides better classification results. In [29], an improved deep residual 3D convolutional neural network (3D-CNN) framework is proposed for detecting and classifying early decay on blueberries. In [30] it is proposed a Spectral-Spatial Network (SSNet) for tobacco impurity detection by integrating HSI and in a [31], the aim was to reconstruct RGB images from HSI and employ deep learning techniques to classify different varieties of corn seeds. Moreover, in [32] HSI is utilized to predict the susceptibility of tomato sepals to fungal infection even before visible symptoms appear using XGBoost and random forest-based regression models.

However, despite its informational richness, leveraging hyperspectral data introduces challenges such as sensitivity to atmospheric conditions (atmospheric haze and instrument noise), illumination changes, and the spectra mixing problem. That imposes limitations for determining the number of pure spectral signatures due to low spatial resolution and heterogenous structures in the image at edges [15]. The high dimensionality of hyperspectral images, while enabling fine spectral discrimination, demands considerable storage and computational power [33]. The similarity and redundancy in neighboring bands present a significant challenge, impacting detection performance when these bands are fed into various modeling frameworks. This underscores the importance of strategic approaches to handling hyperspectral image data effectively [34]. The information redundancy between adjacent spectral bands increases the computational cost, making it challenging to leverage hyperspectral images on a large scale, especially with equipment limitations. Additionally, efforts to eliminate redundant information involve mapping bands into subspaces using a combination of bands that are more informative, less correlated, and more discriminative [34], [35]. Unsupervised ML techniques like PCA and ICA are often employed before applying deep neural network models to hyperspectral images [15], [34], [36]. In studies [34] and [37], authors utilized PCA to extract the representative features of hyperspectral data. In [38] the authors employed the successive projections algorithm, [39], to identify the most important wavelengths and texture features (mean, variance, homogeneity, contrast, dissimilarity, entropy, second moment, and correlation) in HSI. Utilization of typical convolutional neural network (CNN) architecture after reducing the dimensionality of hyperspectral images in the spectrum domain is reported in [40] and [41]. However, a CNN-based method might not perform well in small-sized object detection due to at least two factors: one is the convolutional layer containing a pooling operation, where feature maps may have been down-sampled many times, or spatial information can be lost if the stride is greater than one [41], [42]. Moreover, solutions based on such complex models for object detection in HSI are vulnerable to the peaking paradox [33], [43], [44], [45]. The peaking paradox in the context of hyperspectral images and deep learning establishes that an additional increase in the number of features brings complexity to the classifier as the number of statistical parameters that define the object increases, which leads to an increase in estimation error, negatively affecting the final results [33]. The lack of a large labeled dataset becomes one of the barriers to the use of deep learning in HSI, but strategies like data augmentation offer potential solutions [46], [47], [48]. The study in this paper is a continuation of research from [32], where a semi-automated procedure with limitations in operational usage is developed for the segmentation of the calyx-stem region of tomatoes. The segmentation approach proposed in this study offers a more generalized and robust methodology compared to conventional step-by-step parametric methods that we used in [32], which are often dataset-dependent and lack adaptability to new datasets. While the assessment of susceptibility to fungal infections was demonstrated in our previous study [32], the segmentation method introduced here enhances the extraction of tomato sepals in a more flexible and scalable manner. This approach improves the reproducibility of results and facilitates broader applicability across diverse datasets, thereby strengthening the robustness of the analysis.

Within this paper, an HSI processing pipeline is proposed to perform automatic instance segmentation of the tomato sepal regions based on a regional convolutional neural network (Mask R-CNN) with different backbones. In the context of HSI, where tens of spectral bands are generated, the objective is to encode essential information from the non-noisy components of the spectra and to evaluate how particular encoding affects the performance of the Mask R-CNN instance segmentation model. Among various techniques for encoding spectral information, it is not evident which would be suitable for object detection in hyperspectral data, especially of tomato sepal tips, which consist of a very small number of pixels and therefore require very precise segmentation. Although the review [49] does cover it to some extent, this paper investigates their influence on the successful detection of tomato sepals in hyperspectral images.

II. MATERIALS AND METHODS

A. HSI INSTANCE SEGMENTATION

The field of precision agriculture, with a special focus on post-harvest, is one of the remaining challenging areas where advanced segmentation and detection methods based on deep learning have a great potential to achieve outstanding practical results and thus improve the efficiency of the supply chain and reduce food waste [50], [51], [52]. Researchers continuously explore and propose new architectural designs of deep neural networks (DNN) to address the challenge of accurate classification of the objects within the image by providing pixel-level segmentation and precise estimation of object boundaries [53], [54], [55], [56], [57]. Specifically, CNNs have shown great success in instance segmentation tasks for natural images [58], [59], [60]. The region-based CNN (R-CNN) [61] introduced the concept of region-based object detection, setting the stage for subsequent developments. Fast R-CNN [62], in turn, optimized the speed of this process by introducing a more efficient single-stage detection framework. The MultiPath Network of [63] leveraged multi-scale features to enhance object recognition and segmentation. Faster R-CNN pioneered the use of a Region Proposal Network (RPN) for faster object detection [64]. Mask R-CNN [65] expanded on Faster R-CNN's foundations by introducing pixel-level instance segmentation. Non-local neural networks [66] improved image understanding by capturing long-range feature dependencies. PANet [67] proposed a pyramid attention network for better feature integration. Hybrid Task Cascade by [68] improved performance by cascading multiple tasks, such as object detection and instance segmentation. GCNet [69] incorporated global context information into convolutional networks for better object recognition. Yolact [70] broke ground in real-time instance segmentation using a single-stage detection framework. A Tensor Mask R-CNN model from [71] introduced a tensor-based method for efficient instance segmentation. In the work presented in [72], the authors introduce a transformer-based architecture that directly predicts polygons by using instance Mask R-CNN segmentations as the ground truth supervision for computing the loss. Moreover, the study by [73] proposes a simple and compact ViT architecture called Universal Vision Transformer (UViT), leveraging a constant feature resolution and hidden size throughout the encoder blocks. Through this iterative architecture development process, the DNN models have shown significant performance improvement in instance-based segmentation, their computational efficiency has enhanced, and their domain application has expanded from medical imaging [74] or autonomous driving [75] to precision agriculture [76].

For hyperspectral data, neural networks can be used to tackle tasks such as spectral-spatial feature extraction [77], [78], [79] and segmentation [79], [80], [81] where they are used to identify and segment individual pixels or regions corresponding to specific materials or surfaces in the scene. Recent research studies propose methods for adaptation of deep learning-based instance segmentation models for application on hyperspectral data, such as multi-scale networks [77], [78], [82], spectral-spatial fusion [47], [77], [83], or attention-based models [66], [84], [85]. These methods aim to exploit the spectral and spatial information in hyperspectral data and address the limitations of deep learning models for high-dimensional data. Multi-scale networks are utilized to capture spatial information at different scales, while spectral-spatial fusion methods are used to integrate spectral and spatial information for improved performance on segmentation tasks. The attention-based models weigh the importance of different features or bands for the instance segmentation task.

The existing deep learning-based instance segmentation models are pre-trained on large RGB datasets and can usually be fine-tuned for specific tasks on smaller datasets. These pre-trained models can be further used as a starting point for different tasks without the need to train a deep neural network from scratch, which results in saving time for training and computational resources and usually an improvement in the performance of the final model. In this way, the pre-trained models can be leveraged to apply the knowledge they have learned from natural images and employ it on hyperspectral images. This approach is particularly useful in cases where the number of labeled hyperspectral images is limited.

This study investigates the integration of HSI with deep learning-based instance segmentation through the application of Mask R-CNN architectures. Specifically, three backbone networks—Faster R-CNN, Deformable ConvNet, and FPN—are examined in combination with four spectral encoding techniques- PCA, ICA, PPCA, and NMF- applied to HSIs of the Brioso tomato variety. The objective is to systematically evaluate the impact of these encoding strategies on the performance of instance segmentation networks. To assess the models' generalization capabilities, transfer learning is applied to a secondary dataset comprising three additional tomato varieties: Capricia, Provine, and Sao Paolo. Additionally, the study proposes the design of a custom convolutional layer initialized with spectral encoding coefficients, particularly from NMF, to facilitate an end-to-end, adaptive learning pipeline for robust and accurate tomato sepal instance segmentation in HSI data.

In the following, we reviewed the usually employed dimensionality reduction techniques, putting them in the context of encoding spectral information within HSI.

B. ENCODING HSI SPECTRAL INFORMATION

Consider the hyperspectral image as a 3-order tensor $\mathbf{X} \in \mathbb{R}^{R\times C\times D}$, where R, C corresponds to spatial dimensions (i.e. the image width and height respectively) while D denotes the spectral dimension (i.e. a number of wavelengths). Pixels of the hyperspectral image, associated with the D-dimensional spectral measurements, correspond to the vectors $\mathbf{x}_n \in \mathbb{R}^D, n = 1, \ldots, RC$ within the tensor $\mathbf{X}$. Although the higher values of D, and therefore the number of features in the spectral domain, can have an immense impact on the analysis of the observed phenomena (e.g. plant diseases, fungal infection) [86], [87], this can also pose challenges in computation during hyperspectral image processing. Therefore, finding a suitable trade-off between the computational resources required and the amount of information represented in the spectral domain becomes necessary. The overview [49] of methods used to select the most informative spectral bands and techniques for spectral information encoding, according to hyperspectral image applications, is reported. Within this study, we propose the usage of four techniques for encoding spectral information of hyperspectral images: Principal Component Analysis (PCA) and its probabilistic version (PPCA), Independent Component Analysis (ICA), and Non-negative Matrix Factorization (NMF). All four techniques find common applications as initial preprocessing steps in hyperspectral image analysis, serving multiple purposes such as dimensionality reduction, encoding spectral information, signal-to-noise ratio enhancement, and the mitigation of complexity of computation during subsequent data analysis and machine learning tasks [49].

PCA, also known as the Karhunen-Loève transform, is a commonly used technique for dimensionality reduction or feature extraction [88]. According to [89], PCA can be defined as the orthogonal linear projection of the data onto a lower-dimensional, so-called principal subspace, which maximizes the variance within it. Consider a data set of hyperspectral responses per pixel $\{\mathbf{x}_n \in \mathbb{R}^D | n = 1, 2, \ldots, N\}$ where $N = RCL$, (i.e., RC is the image width and height, respectively, and L is the number of images). Denoting with $\bar{\mathbf{x}}$ the sample mean of the considered set, a sample covariance matrix $\mathbf{M} \in \mathbb{R}^{D\times D}$ is computed as

$$\mathbf{M} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^T. \tag{1}$$

From the factorization of the $\mathbf{M}$ using eigendecomposition:

$$\mathbf{M} = \mathbf{U}3\mathbf{U}^T \tag{2}$$

where $\mathbf{U}$ is an orthogonal matrix in $\mathbb{R}^{D\times D}$, whose columns are the eigenvectors of $\mathbf{M}$, and $3 \in \mathbb{R}^{D\times D}$ is a diagonal matrix where the corresponding eigenvalues are on diagonal. By selecting only the $K < D$ dominant eigenvectors (i.e., principal components) $\mathbf{u}_1, \ldots \mathbf{u}_K$ that are associated with the largest eigenvalues $\{\lambda_1, \ldots, \lambda_K\}$ on 3's diagonal, a truncated orthogonal projection matrix $\mathbf{U}_K$ is formed. Then the spectral information in $\mathbf{x}_n$ is encoded with K principal components as:

$$\mathbf{z}_n = \mathbf{U}_K^T(\mathbf{x}_n - \bar{\mathbf{x}}) \tag{3}$$

where $\mathbf{z}_n \in \mathbb{R}^K$ is a representation of $\mathbf{x}_n$ within the K-dimensional principal subspace.

The reformulation of the PCA, as a probabilistic model (i.e. PPCA) was proposed in [90] and was closely related to factor analysis [91], [92]. Within PPCA, an explicit latent variable $\mathbf{z}_n \in \mathbb{R}^K$, corresponding to the principal subspace spanned by the columns of $\mathbf{W} \in \mathbb{R}^{D \times K}$, is introduced and refers to the observed variable $\mathbf{x} \in \mathbb{R}^D$ through linear transformation:

$$\mathbf{x} = \mathbf{W}\mathbf{z} + \varepsilon \tag{4}$$

where $\varepsilon \in \mathbb{R}^D$ and $\varepsilon \sim \mathcal{N}(\mu, \sigma^2\mathbf{I})$. Given that the prior distribution is $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ and the conditional distribution is $p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{W}\mathbf{z} + \mu, \sigma^2\mathbf{I})$, the linear-Gaussian model from equation 4 and using Bayes' rule, the posterior distribution is

$$p(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{C}^{-1}\mathbf{W}^T(\mathbf{x} - \mu), \sigma^2\mathbf{C}^{-1}),$$

where

$$\mathbf{C} = \mathbf{W}^T\mathbf{W} + \sigma^2\mathbf{I}.$$

[90], [93]. With the maximum likelihood estimation of $p(\mathbf{z}|\mathbf{x})$ parameters denoted as $\mathbf{W}_{ML}, \mu_{ML}, \sigma_{ML}$ any latent representation of $\mathbf{x}$ is summarized through the mean of $p(\mathbf{z}|\mathbf{x})$ as:

$$\mathbb{E}_{p(\mathbf{z}|\mathbf{x})}[\mathbf{z}] = \mathbf{C}^{-1}\mathbf{W}_{ML}^T(\mathbf{x} - \mu_{ML}). \tag{5}$$

with $\mu_{ML} = \bar{\mathbf{x}}$ and

$$\mathbf{W}_{ML} = \mathbf{U}_K^T(3_K - \sigma^2\mathbf{I})^{1/2}\mathbf{R}. \tag{6}$$

where $\mathbf{R} \in \mathbb{R}^{K \times K}$ is an arbitrary orthogonal rotation matrix and matrices $\mathbf{U}_K \in \mathbb{R}^{D \times K}, 3_K \in \mathbb{R}^{K \times K}$ are truncated version of matrices $\mathbf{U}, 3$ obtained by eigendecomposition of covariance matrix $\mathbf{M}$ from equation 2. In the limit when $\sigma^2 \to 0$, $\mathbf{C}^{-1} \to (\mathbf{W}_{ML}^T\mathbf{W}_{ML})^{-1}$ equation 5 represents an orthogonal projection into the latent principal subspace and thus standard PCA spectral information encoding is performed.

Unlike PCA or its probabilistic counterpart PPCA, which focuses on maximizing variance, ICA aims to find statistically independent components in the hyperspectral signature, which can be particularly useful for separating spectral sources corresponding to different materials [94], [95], [96]. Consider the data matrix $\mathbf{Y} \in \mathbb{R}^{D \times N}$ obtained by centering columns of the data matrix $\mathbf{X} \in \mathbb{R}^{D \times N}$ by equation:

$$\mathbf{y}_n = \mathbf{x}_n - \bar{\mathbf{x}}. \tag{7}$$

The ICA technique models the centered data matrix $\mathbf{Y}$ as:

$$\mathbf{Y} = \mathbf{A}\mathbf{S}, \tag{8}$$

where $\mathbf{A} \in \mathbb{R}^{D \times D}$ represents the mixing matrix for linear combinations of independent components, while $\mathbf{S} \in \mathbb{R}^{D \times N}$ is the source matrix, with the independent components as its columns. Then the goal of ICA is to find an orthogonal unmixing matrix $\mathbf{Q} \in \mathbb{R}^{D \times D}$ such that:

$$\mathbf{S} = \mathbf{Q}\mathbf{Y}. \tag{9}$$

With the application of the unmixing matrix to the centered data $\mathbf{y}_n$, the spectral information is encoded with K-independent components:

$$\mathbf{s}_n = \mathbf{Q}_K\mathbf{y}_n, \tag{10}$$

where $\mathbf{s}_n \in \mathbb{R}^K$ represents the $\mathbf{y}_n$ within the K-dimensional subspace, obtained by linear transformation $\mathbf{Q}_K \in \mathbb{R}^{K \times D}$.

The last technique we employed in this study for encoding spectral information, NMF, aims to factorize the data matrix $\mathbf{X}^{D \times N}$ into two non-negative matrices, $\mathbf{V} \in \mathbb{R}^{D \times K}$ and $\mathbf{H} \in \mathbb{R}^{K \times N}$:

$$\mathbf{X} = \mathbf{V}\mathbf{H}, \tag{11}$$

where K is the desired reduced dimensionality of the subspace where spectral information is encoded. An essential aspect of NMF is that both $\mathbf{V}$ and $\mathbf{H}$ are constrained to contain non-negative elements. This non-negativity constraint helps reveal additive components within the data. This approach involves iterative optimization techniques to minimize the reconstruction error between the original data $\mathbf{X}$ and its factorization $\mathbf{V}\mathbf{H}$. Once the factorization is obtained, the data $\mathbf{x}_n$ can be represented in the reduced K-dimensional space as:

$$\mathbf{z}_n = \mathbf{H}^T\mathbf{x}_n, \tag{12}$$

where $\mathbf{z}_n \in \mathbb{R}^K$ captures the representation of $\mathbf{x}_n$ within the K-dimensional NMF subspace.

C. INSTANCE SEGMENTATION OF TOMATO SEPALS

In this study, we dealt with the problem of adapting a hyperspectral image for training and testing existing DNN instance segmentation models on a tomato sepal detection task. Although the main advantage of this type of data is its richness in spectral information, offering wide application directions, it still requires high computational costs. Advanced instance segmentation DNN models are usually pre-trained on RGB images [97], [98], [99]. Utilization of the transfer learning approach on pre-trained DNN instance segmentation models to detect tomato sepals in hyperspectral images requires the compression of image spectral information from a high-dimensional to at least a three-dimensional latent space. Specifically, within this study, our methodology involves employing linear techniques from Subsection II-B to effectively encode the spectral information within noise-reduced hyperspectral images by preserving the most significant spectral information for further fine-tuning pre-trained instance segmentation DNN-based models by using a transfer learning approach.

This approach exemplifies the broader concept of transfer learning in machine learning, which has gained popularity due to its potential applications [100]. Transfer learning involves improving learning in a new task by transferring knowledge from a related task that has already been learned to improve performance in the target task, as presented in this study. The principle of transfer learning often involves mapping characteristics from one task to another to establish correspondences, ultimately aiming to achieve higher initial performance, a steeper learning curve, or a higher final performance level through transfer learning by avoiding negative transfer (this occurs when knowledge learned from one task hinders or degrades the performance of a model on a different but related task) and automating the mapping process [101]. By doing so, transfer learning can overcome the challenges of limited data, reduce training time, and improve the performance of machine learning models in various domains.

Based on the result of applied linear techniques for encoding spectral information, we initiated fine-tuning of pre-trained deep learning models on encoded hyperspectral images to detect and segment tomato sepals. In this work, we focus on Mask R-CNN [65], which is a two-stage and well-known architecture for instance segmentation tasks. It builds upon the Faster R-CNN framework by adding an additional branch for generating object Mask R-CNNs alongside bounding box predictions [64]. Hence, Mask R-CNN detects and segments multiple objects within an image simultaneously. The architecture of Mask R-CNN consists of two main components: a backbone network and two task-specific subnetworks. The backbone network is based on a pre-trained CNN such as ResNet [102] to extract high-level features from the input image, which further propagate through two parallel subnetworks: the region proposal network (RPN), which generates potential object proposals by predicting bounding boxes, and the second, which takes the proposed regions from the RPN and refines their bounding box coordinates while simultaneously predicting the class probabilities for each object category. In addition, it generates a binary mask for each proposed region, delineating the objects at the pixel level. Since the output of Mask R-CNN contains an object mask, it is a reasonable and appropriate choice for tasks that require precise object localization and segmentation, such as tomato sepals.

Another advantage of Mask R-CNN is its facility for transfer learning and its adaptability to new problems [103]. The Mask R-CNN models used in this study are pre-trained on Microsoft's Common Objects in Context, or COCO dataset, a large-scale dataset with 200k images [98]. Three types of backbones are used: FPN (ResNet + FPN), C4 (ResNet) [64] and DC5 or Dilated-C5 (ResNet with dilations in the conv5 backbone) [104]. The selection of these backbones was guided by their architectural features: FPN was chosen for its multi-scale feature extraction, C4 (ResNet) for its robust performance, and DC5 (Dilated-C5) for its suitability for tasks demanding dense predictions and precise spatial information. The proposed pipeline for detection of tomato sepals and segmentation of their regions within HSI is depicted in Figure 1, starting with encoding spectral information, and then across fine-tuning Mask R-CNN models with three different backbones.

D. DATASETS

This study employs two datasets containing hyperspectral images of different tomato varieties. Hyperspectral images are acquired using the SPECIM FX17 camera, which measures reflectance within the spectral range from 900 nm to 1700 nm with a resolution step of 3.46 nm thus providing 224 spectral bands per image. To mitigate the influence of lighting conditions on hyperspectral imaging, all image acquisitions were conducted in a controlled darkroom environment, utilizing halogen lamps as the sole illumination source. This setup ensured uniform lighting conditions and eliminated potential interference from external light sources, thereby enhancing the reliability of the spectral data. This ensured that the hyperspectral dataset remained unaffected by external illumination inconsistencies while minimizing redundant spectral information, thereby improving the reliability and reproducibility of the study. The dataset acquisition procedure was performed at Wageningen Food and Biobased Research (WFBR) within Wageningen University and Research.

The first dataset contains hyperspectral images of the Brioso cultivar (6 batches) formed after harvest by five different growers based in the Netherlands and Belgium. After being harvested in greenhouses without supplementary lighting, tomatoes were delivered the same day for imaging and stored at a temperature of 15°C. Tomatoes were pruned and distributed into trusses, containing three or four tomatoes, which were then imaged. Having six trusses per batch resulted in 36 hyperspectral images of tomato trusses, with an average number of 19 tomato sepals per image (approximately 5 sepals per tomato) [32] (see Table 1).

The second dataset was acquired under a distinct experimental protocol, involving a reduced count of tomatoes (one tomato as opposed to three or four from the first dataset image acquisition). Furthermore, a discrepancy in the sensor-to-object distance during data collection led to decreased resolution (i.e. number of pixels per object of interest such as tomato sepals) for the second dataset. This discrepancy arose due to the tomatoes being positioned at a greater distance from the sensor, ultimately contributing to the observed disparity in resolution between the two datasets. The tomatoes of four varieties: Brioso, Capricia, Provine, and Sao Paolo were used during the creation of the second dataset. Tomatoes were produced in a greenhouse in De Lier, the Netherlands, and supplied by Growers United at WFBR. Among varieties, there are noticeable differences in size. Brioso is medium to large, as is Provine, while Capricia and Sao Paolo are generally smaller in size, often referred to as cherry tomatoes. Moreover, Sao Paolo has a distinctive deep red color. The second dataset contains two hyperspectral images containing 32 tomatoes for Brioso and Sao Paolo varieties, and two hyperspectral images with 16 tomatoes each per Capricia and Provine varieties (see Table 1).

For achieving similar scales of tomatoes, thus sepals between datasets, hyperspectral images from the second

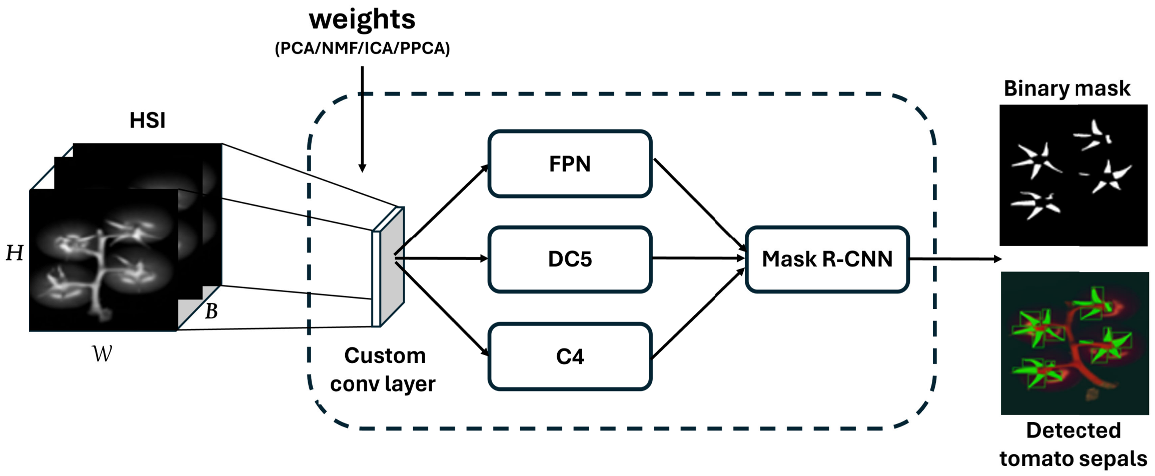

FIGURE 1. The workflow of the HSI data processing pipeline.

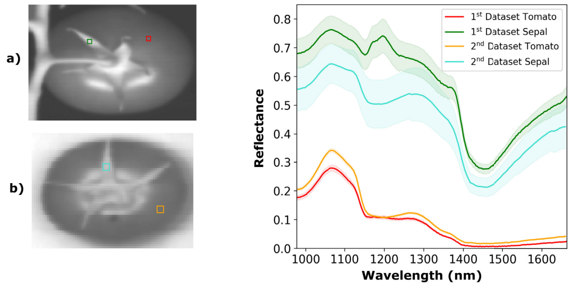

FIGURE 2. Hyperspectral responses of tomato fruit and tomato sepal on examples from the a) first and the b) second dataset.

TABLE 1. Datasets description.

	Variety	Images	Sepals
Dataset 1	Brioso	36	698
Dataset 2	Brioso Capricia Provine Sao Paolo	4×32	164 167 163 158

dataset are divided such that there is one tomato per image and the image is resized to correspond to the size of the region which covers one tomato in the image in the first dataset. Fig 2 shows examples of spectral responses of tomato fruit and tomato sepals for the same variety Brioso in both datasets. Hyperspectral responses confirmed the initial hypothesis that the reflectances from the tomato fruit and tomato sepals in the observed range correlate among the two datasets. The correlation among the tomato sepals is 0.8859 with p-value $0.87 * 10^{-35}$, and among tomato fruits is 0.8953 with p-value $1.75 * 10^{-35}$. Figure 3 illustrates the variability in tomato sepal width (50-125 pixels), height (35-120 pixels), and area (500-4200 pixels), providing crucial information on the morphological diversity within the second dataset and among datasets.

The preprocessing steps for hyperspectral images that consist of black-and-white correction followed by noise reduction per spectral band are adopted from [32] where they are briefly discussed. After the noise reduction step, the first thirteen and last sixteen spectral bands from initial 224 are removed due to low signal-to-noise ratio [32] leaving 196 spectral bands for further analysis. These bands are then subjected to standard normal variate (SNV) spectral correction technique. By applying SNV correction, we got enhanced spectral feature extraction, reduced effects of unwanted variations, such as baseline shifts and scaling, and improved comparability between spectra [105].

The hyperspectral images are additionally divided such that newly created images contain only one tomato per image. All ground truth polygon-wised masks for the detection of tomato sepals are created using LabelMe software [106]. Groundtruth for counting tomato sepals is obtained by numbering the marked tomato sepals from JSON files made in LabelMe software. The methodology is implemented in Python v3.9 using the following open-source libraries: Spectral Python (SPy) [107] and Scikit-Image (skimage) [108] for data preprocessing, Scikit-Learn (scikit) for the implementation of spectral information encoding techniques [109], and detectron2 [110] for employing DNN instance segmentation models for tomato sepal detection.

E. HARDWARE

In this study, complex computational operations were executed using a single high-performance graphics processing unit (GPU). The GPU employed for deep learning task was the NVIDIA GeForce RTX 2080 Ti. With 4352 CUDA cores and 11 gigabytes of high-bandwidth memory (HBM2), it provides the necessary computational resources to expedite the training and inference phases of considered deep neural network models.

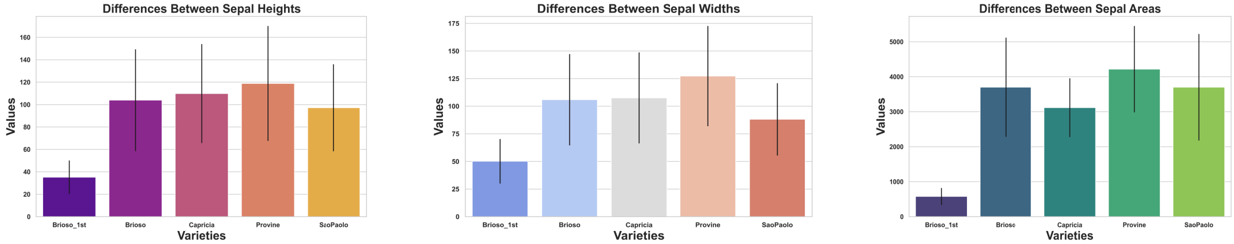

FIGURE 3. The diversity in heights, widths, and areas of tomato sepals in both datasets across varieties.

F. EVALUATION METRICS

In this study, the mean average precision (mAP) metric is used for the performance evaluation of instance segmentation models. Since we consider the single-class case problem [111], [112], in the following, with mAP we denote mAP_{50} where the adopted threshold for IoU calculation was set to 50% of the overlap among a ground truth and predicted region which corresponds to tomato sepal (i.e. mAP - 50 will be denoted as mAP in the following text). Intending to estimate the precision in the number of detected sepals, we highlight two error rates: under- and over-predictions, which give us a better insight into the performances of the models [113], [114]. Estimating the type of errors is important for future operational usage (e.g., logistic and transport optimization, assessment of shelf life) where decision-making would be performed by aggregating sepal level predictions at individual tomato or tomato truss level. The over-prediction error pertains to the count of identified sepals in a test image not present in the ground truth data. In other words, it signifies instances where the algorithm falsely detects sepals that are not genuinely part of the tomato (false positive). Conversely, the under-prediction errors refer to cases where the algorithm fails to detect tomato sepals that are verified as part of the ground truth data for a given test image. The motivation for using these two metrics comes from the fact that when the simple mean absolute error is used, these values (missed or false detections) can cancel each other out through averaging. Through the following equations:

$$\overline{\mathbf{e}} = \frac{\mathbf{FP}}{\mathbf{P}} \quad , \quad \underline{\mathbf{e}} = \frac{\mathbf{FN}}{\mathbf{P}} \tag{13}$$

the errors of over $\overline{\mathbf{e}}$ or under $\underline{\mathbf{e}}$ predicted tomato sepals are calculated, where **P** is the number of sepals within ground truth, **FP** is the number of falsely detected sepals, and **FN** denotes the number of non-detected sepals.

III. RESULTS

A. EVALUATION OF THE PROPOSED METHODOLOGY

A detailed analysis of how the choice of technique for encoding spectral information of hyperspectral images in conjunction with one of the deep learning-based instance segmentation models affects the final identification accuracy of tomato sepals is presented below. The PCA often focuses on detecting linear correlations among variables, which may not be ideal for capturing complex relationships present among hyperspectral bands, and it might face limitations in scenarios where mean and covariance statistics alone fail to adequately represent the data's intricate structure. The PPCA, a probabilistic extension of PCA, can capture the underlying probabilistic structure of data, which may be particularly relevant when dealing with uncertainty in hyperspectral information. The ICA, on the other hand, aims to discover statistically independent sources within the data, which can unveil underlying physical or chemical properties in hyperspectral imagery [115]. The NMF, with its non-negative constraints, can reveal additive components and part-based representations in the data, which is valuable for interpreting complex and detailed spectral information esspecially in segmentation domain, because objects in images are often composed of additive parts, and negative values can be meaningless in many physical contexts (e.g., spectral reflectance or pixel intensities) This characteristics make NMF advantageous for tasks like hyperspectral image segmentation, where pixel intensities (and their derived features) must be non-negative. This constraint ensures that the extracted features (components) are more interpretable, as they represent additive combinations of the original features, which is essential when analyzing biological structures like tomato sepals. Additionally, NMF focuses on extracting meaningful parts-based representations of the data, which aligns well with the goal of segmenting tomato sepals. In contrast, PCA and ICA focus on maximizing variance and statistical independence, respectively. However, they may not always lead to parts-based features that are directly interpretable or well-suited for segmentation tasks. The ability of NMF to discover components that represent additive parts makes it a more effective method for isolating distinct regions or features in hyperspectral images. While PCA and ICA tend to produce components with lower variability (as indicated by their lower standard deviations), these components are often more abstract and

TABLE 2. For each technique for encoding spectral information, the sum of standard deviations for all coefficients, estimated for projection onto a three-dimensional subspace, across all folds is reported.

PCA	PPCA	ICA	NMF
0.38	2.87	10.21	13.52

All values in table are multiplied with 10^{-4}

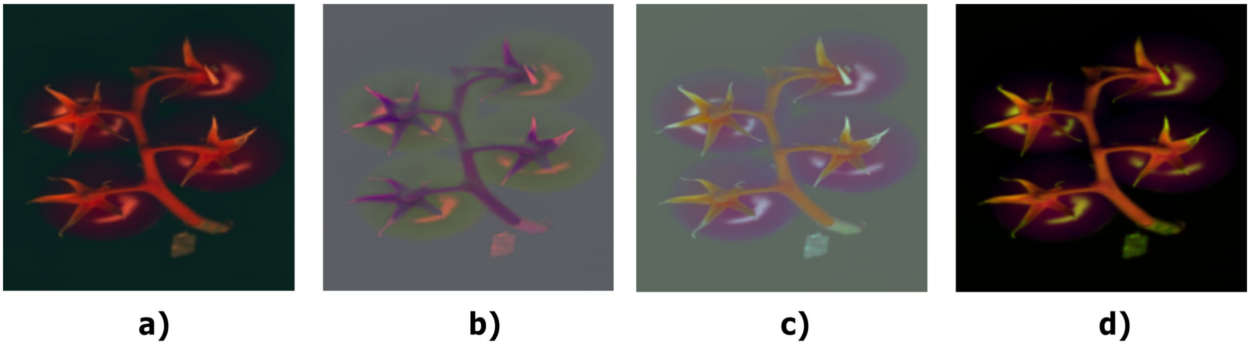

FIGURE 4. Encoded spectral information of one image from dataset 1 within a three-dimensional subspace using: a) PCA, b) ICA, c) PPCA, and d) NMF techniques represented through a three-channel false-color image.

less interpretable. In contrast, NMF's components, while showing slightly higher variability than PCA and ICA, are still relatively stable (as evidenced by the smaller standard deviations compared to FA). This suggests that NMF strikes a good balance between feature extraction and stability. Moreover, as supported in the literature [116], NMF provides a robust framework for hyperspectral unmixing due to its non-negativity constraint and the interpretability of the resulting components. Additionally, hyperspectral image segmentation was successfully accomplished, after NMF in [117], as in our study. Therefore, the choice between the mentioned techniques for hyperspectral data analysis should be guided by the specific goals, characteristics, and complexities of the dataset and the intended downstream tasks, acknowledging the distinct advantages and limitations of each method.

For this purpose, we used Dataset 1 from Table 1 in which 36 images are distributed in training, validation, and test set in a ratio: 84%, 8%, and 8% respectively (30-3-3). The projection for each technique for encoding spectral information is learned by running a 6-fold cross-validation with the adopted dataset division ratio.

We have included a summary of the sum of standard deviations for all coefficients, estimated for the projection onto a three-dimensional subspace, across all folds (see Table 2). This provides a quantitative comparison of the variability across the different encoding techniques (PCA, PPCA, ICA, and NMF) and further supports the rationale for choosing NMF for tomato sepal segmentation. While NMF shows the highest sum of standard deviations across all folds, this is not necessarily a disadvantage. The increased variability reflects the method's ability to capture more complex features that are well-suited for the segmentation task. NMF's non-negative constraint ensures that these components are interpretable and correspond to real-world regions in the image, which is crucial for tasks like identifying distinct parts of tomato sepals. By comparing to the with other techniques, PCA and PPCA have smaller sums of standard deviations, indicating that their components are more stable but less complex and less informative in the context of hyperspectral data. They are good for dimensionality reduction but might not provide the detailed, parts-based features necessary for accurate segmentation. ICA captures independent components but introduces larger variability, making it less reliable for segmentation, especially for tasks requiring consistent and interpretable features. We found that differences in learned projections among the folds were not statistically significant for each technique (see Table 2). Since the variations between the folds were observed only at the level of the fourth decimal point, we randomly selected one of the possible six learned transformations per technique and used it in future experiments. Examples of a transformed hyperspectral image from Dataset 1 with learned projections are given in Figure 4.

TABLE 3. The mean and standard deviation of mAP across 6 validation folds for three Mask R-CNN model variants (FPN, C4, and DC5) and used techniques for encoding spectral information (PCA, NMF, ICA, and PPCA).

	PCA	PPCA	ICA	NMF
C4	92.92 ± 3.43	93.50 ± 4.12	93.33 ± 3.03	92.19 ± 3.84
DC5	91.45 ± 3.17	93.54 ± 3.23	93.21 ± 2.71	92.09 ± 3.39
FPN	92.39 ± 3.33	92.93 ± 3.39	92.85 ± 4.22	**94.05 ± 3.25**

The proposed pre-trained instance segmentation models [64], [65], [104] (see Section II-C) are further fine-tuned on hyperspectral images in which noise-reduced 196-dimensional spectral information is encoded within a three-dimensional subspace with techniques from Section II-B). The augmentation applied during inference involves resizing the input image while preserving its aspect ratio. This process is governed based on dataset statistics related to the desired minimum size of the shorter edge and the maximum allowable size of the longer edge. This essentially means that the image's specific size is randomly selected from within the specified range. This augmentation guarantees that the image is resized to have a minimum shorter edge while maintaining its aspect ratio. Nevertheless, this type of augmentation induces modifications through transformations applied to the images, giving diverse representations to provide variety and improve the model's robustness. The changes made are based on resizing and flipping operations. We chose these techniques to address the challenge of varying tomato sepal positions within the image. The flip technique helps simulate different sepal orientations, ensuring that the model can recognize sepals in various orientations, while resize helps to standardize image sizes, making the sepal's features more consistent for analysis. Together, these simple augmentation techniques allow better handling of diverse sepal directions, improving the model's ability to accurately assess them regardless of their position in the image.

Through the 6-fold cross-validation training process, we employed one image per batch (each containing around 20 sepals or instances) to optimize memory efficiency when transferring data to and from the GPU for both datasets separately and merged. This strategy aimed to improve generalization performance on new data [118], [119]. Additionally, we used a warm-up cosine learning rate scheduling [120] that gradually increased the learning rate from a base value (0.001) to its target value a priori, which then subsequently decreased over a specified number of epochs (5000), following a cosine curve. This approach facilitates gentle parameter space exploration, averts significant updates that might disrupt the optimization trajectory within parameter space, and fosters smooth adjustments during fine-tuning and convergence [120], [121], [122].

Table 3 presents the performances of the proposed Mask R-CNN model with different backbones in terms of mAP, with considered four techniques for spectral information encoding. The overall assessment of the model's performance encompasses a dual perspective: firstly, through the utilization of the mAP metric to evaluate the success of instance segmentation of tomato sepals, and secondly, by considering detecting precision. The primary metric is the mean average precision, where the NMF technique in conjunction with the Mask R-CNN + FPN backbone (in further text denoted as NMFMaskFPN) achieves the best results, reaching a mAP of 94.05 ± 3.25%. Figure 5 presents the results obtained by the Mask R-CNN + FPN model with all four techniques for spectral information encoding, accompanied by the obtained binary masks and predictions of the tomato sepal identification. In addition to the mAP, further insights into model efficiency were gleaned from three supplementary metrics: accuracy, false negative rate, and false positive rate through training epochs. The average accuracy and sensitivity values for the evaluated models (FPN, DC5, and C4) exhibit minimal variation, with the average accuracy values being 0.9288, 0.9256, and 0.9135. The average sensitivity values are 0.8821, 0.8788, and 0.8695, respectively, indicating a high degree of similarity in their performance. Besides these metrics, the false negative and the false positive rates also exhibit comparable performance across the used Mask R-CNN configurations on hyperspectral data images encoded by the NMF technique (see Figure 6); even marginal enhancements in these metrics could yield discernible improvements in the precision of tomato sepal detection outcomes.

Each of the four techniques for encoding spectral information addresses the presence of specular reflection differently. Notably, seen in Figure 4, both sepal tips and regions under specular reflection exhibit similar responses after encoding spectral information. This can lead to potential misclassification of sepals. Since the primary objective of this study is to accurately detect and segment the region of the sepals, for further use as input to a prediction model from [32], and for evaluation of the accuracy of the model in sepal detection, an approach with two-sided perspective is employed. This approach considers errors in both under-predicted and over-predicted numbers of tomato sepals, which are briefly explained in the section II-F. The study's findings are presented in Table 4, where the form $\overline{\mathbf{e}}/\underline{\mathbf{e}}$ is used to denote errors in the prediction of tomato sepals on the test set. The NMFMaskFPN approach stands out as it achieves zero under-predictions across all categories, demonstrating better prediction performance on the test set in terms of both over- and under-predictions compared to other backbones, attaining a perfect 100% accuracy in detecting sepals.

TABLE 4. The $\overline{\text{e}}/\underline{\text{e}}$ prediction errors of tomato sepals on the test set expressed in percentage.

	PCA	PPCA	ICA	NMF
C4	1.69/0	1.69/0	1.69/1.69	3.39/0
DC5	0/1.69	0/1.69	1.69/0	0/1.69
FPN	0/1.69	1.75/1.69	1.75/3.39	**0/0**

Further, we evaluate NMFMaskFPN trained on hyperspectral images of Brioso variety i.e. Dataset 1, on hyperspectral images of the same variety from Dataset 2. The hyperspectral images from Dataset 2 are prepared by following the same procedure as for Dataset 1. Even though the obtained mAP on the second dataset is very low, the model detected 37% of the total number of sepals (61/164 on 32 images), from which 77% are true positive and 0.29% are false positive, including 2 sepals with double detection. The poor metrics during validation on Dataset 2 could be caused by the distribution shift between the two datasets. In Dataset 1, the model was trained to recognize tomato sepals in a truss of four tomatoes, where it learned to detect more complex shapes and relationships between multiple fruits and sepals. In contrast, Dataset 2 contains images with only one tomato containing a few sepals, which is significantly different and may cause poor performance. Other factors, such as different contexts and scale differences (where the individual sepal is likely larger compared to a sepal in a truss) and the model's focus on recognizing tomato sepals in a truss rather than individual sepals, can all contribute to this issue. Additionally, if the model is overfitted to specific characteristics, it might struggle to generalize to isolated tomato fruit images. To improve performance, we consider fine-tuning the model on Dataset 2 and employing transfer learning to adapt the model trained on Dataset 1.

We conducted further training of the NMFMaskFPN model through 8-fold cross-validation, using 86% of data from Dataset 2 for training (28 of 32 (28/32) images for each variety which is 130 tomato sepals per variety, or all varieties merged together: 112/128 images which is 570 sepals), and with the 7% data (2/23 images for each variety which is 16 tomato sepals or 8/112 images of all varieties which is 41 tomato sepals) for each validation and testing set. We follow the same training procedure conducted with

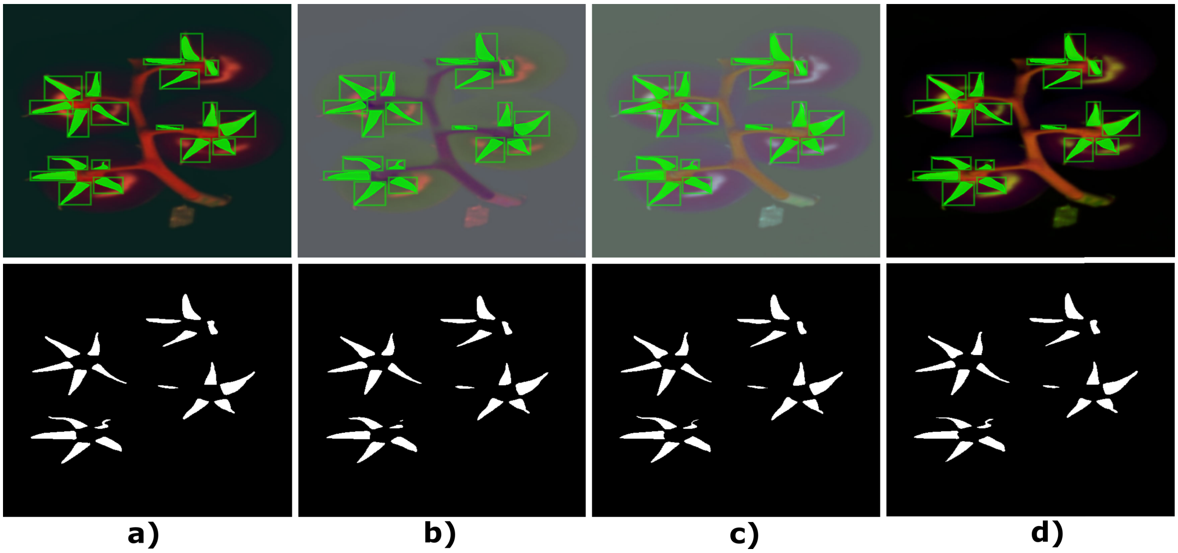

FIGURE 5. Detected instances of tomato sepals by NMFMaskFPN model utilizing: a) PCA b) ICA c) PPCA d) NMF techniques for spectral information encoding (the first row), and belonging binary masks (the second row).

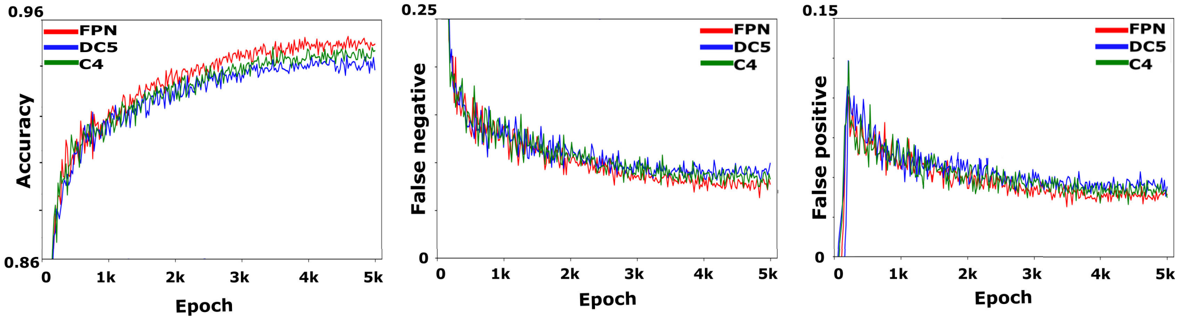

FIGURE 6. Performance analysis of Mask R-CNN models through epochs with different backbones (FPN, DC5, and C4) evaluated on encoded hyperspectral images from the validation set by the NMF technique.

Dataset 1, containing only one variety (Brioso), now for four varieties individually from Dataset 2, and all four varieties merged from Dataset 2. The performances of the NMFMaskFPN model after additional training are shown in Table 5.

TABLE 5. The mean and standard deviation of mAP for NMFMaskFPN across 8-fold-cross validation.

Brioso	Capricia	Provine	Sao Paolo	Merged
94.0 ± 8.4	97.4 ± 7.4	98.5 ± 4.3	82.2 ± 12.3	86.4 ± 7.2

The presented results demonstrated that the proposed transfer learning approach achieves better performance and shows robustness on the altered experimental design and varied instances per image, showing the improved generalization among different varieties (see Figure 7).

The proposed approach showed the best results on the test dataset for the Provine variety. In contrast, the poorest results were obtained for Sao Paolo, attributable to the very small size of the fruit, tomato sepals, and lower resolution. Despite lower precision in segmentation for some varieties the final accuracy in detecting sepals is not affected. The model from Table 5 demonstrated efficiency in the detecting of tomato sepals within the second dataset for Brioso, Provine, and Sao Paolo varieties, achieving 98.52% precision on the test set. For Capricia, the model exhibited a single false-positive detection on the test set.

TABLE 6. The mAP obtained through pre-initialization of ConvMaskFPN using coefficients derived from the following techniques: XavierN, KaimingN, Uniform, and normal.

XavierN	KaimingN	Uniform	Normal
82.24 ± 4.89	82.18 ± 4.49	81.12 ± 4.52	82.07 ± 3.13

We further explore the performance of the Mask R-CNN model by adding a convolutional layer at the beginning of the architecture (named ConvMaskFPN in further text) for encoding spectral information of hyperspectral images. We first, investigate various weight initializations of this added layer using four initialization techniques: XavierNormal (XavierN), KaimingNormal (KaimingN), Uniform, and Normal [123]. The results (see Table 6) on a merged dataset are obtained by fine-tuning the baseline model, Mask R-CNN with FPN backbone, with an additional 5k epochs. We can see that the performances of ConvMaskFPN are similar to those

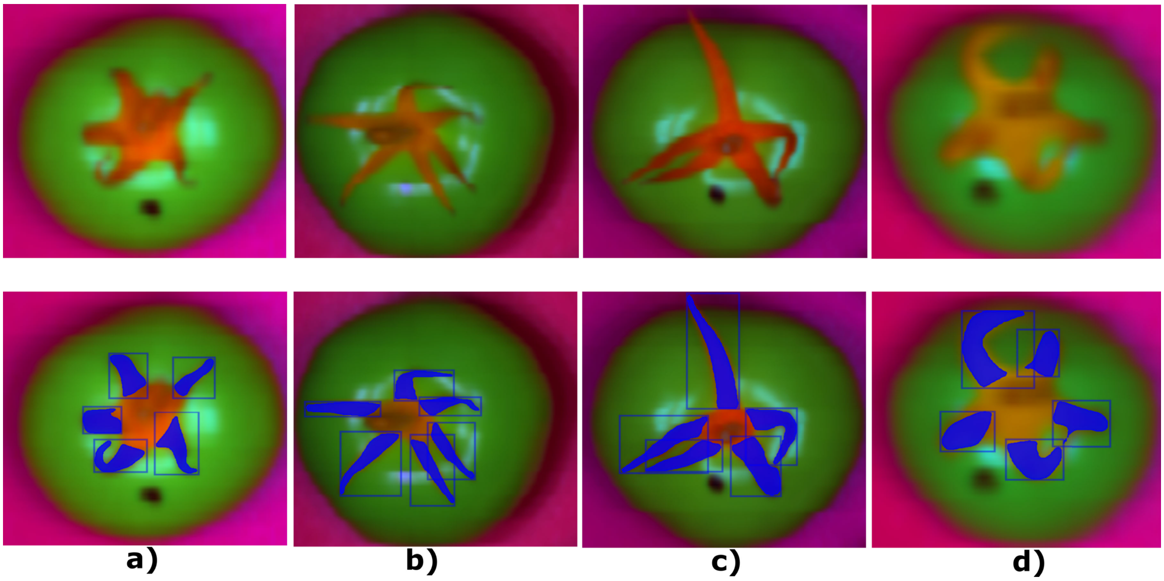

FIGURE 7. Results of tomato sepals detection on four cultivars from the dataset 2: a) Brioso b) Capricia c) Provine d) Sao Paolo with NMFMaskFPN.

of XavierN and KaimingN and slightly better than those with Uniform and Normal weight initialization.

TABLE 7. The mAP obtained with pre-initialization of ConvMaskFPN using coefficients derived from used spectral information encoding techniques.

PCA	PPCA	ICA	NMF
86.95 ± 3.63	87.35 ± 2.44	86.96 ± 2.01	**87.40 ± 3.81**

Moreover, we initialized weights of the added layer with learned coefficients from the spectral information encoding techniques. Further, we fine-tuned the entire ConvMaskFPN, evaluated the performance on the validation and test datasets, and compared it with the performance of the baseline Mask R-CNN model, NMFMaskFPN. Our findings revealed a slight improvement in overall precision, giving more stable predictions. On the merged dataset comprising diverse varieties, the addition of the extra layer led to a notable 1% enhancement in precision with NMF coefficients as initialization (see Table 5 and Table 7). The final model demonstrated a heightened ability to comprehend the NMF representation of hyperspectral images, resulting in more robust detections and enhanced precision. This observation underscores the substantial potential of leveraging the Mask R-CNN framework pre-initialization with NMF coefficients, which contributes to improved performance in tomato sepal detection.

IV. DISCUSSION

Within the framework of the presented study on the segmentation of tomato sepal instances, the performance of the Mask R-CNN model with different backbones (C4, DC5, and FPN) and different spectral information encoding techniques was compared. The Mask R-CNN model with an FPN backbone and integrated with NMF as an adopted spectral information encoding technique for hyperspectral data transformation, named as NMFMaskFPN, achieved the highest mAP and demonstrated superior stability, making it a reliable choice for the instance segmentation of tomato sepals.

Additionally, considering the variation in over- and under-prediction of tomato sepals for different models and spectral information encoding techniques suggests that both choices can significantly impact the accuracy of results depending on specific tomato structures. This highlights the importance of adapting pre-trained models to new datasets, especially when the conditions during image acquisition vary. Fine-tuning the selected NMFMaskFPN model on the target dataset helped to optimize its performance and improve its generalization, demonstrating the robustness of the proposed approach across different tomato varieties. Moreover, the incorporation of an extra custom layer through the ConvMaskFPN model and initializing it with NMF coefficients, yielded a modest yet discernible enhancement in overall precision of 1%, reaching the highest mAP of 87.4%, fostering more reliable predictions.

The differences in simple morphometric characteristics of sepals across tomato varieties (see Figure 3) expressed through the number of pixels indicate the difficulty of the precise sepal segmentation problem. The percentage of border pixels within the tomato sepals directly influences the mean and standard deviation of the hyperspectral response per sepal, and incorrect segmentation of the border pixels can drastically change this response and potentially lead to the misclassification of infected sepals [32]. The segmentation masks generated in this work are intended to facilitate future investigations by ensuring that each pixel is accurately segmented, thereby preserving critical structural details of the sepals. This is particularly important for the tips of the sepals, which are exceptionally thin and highly

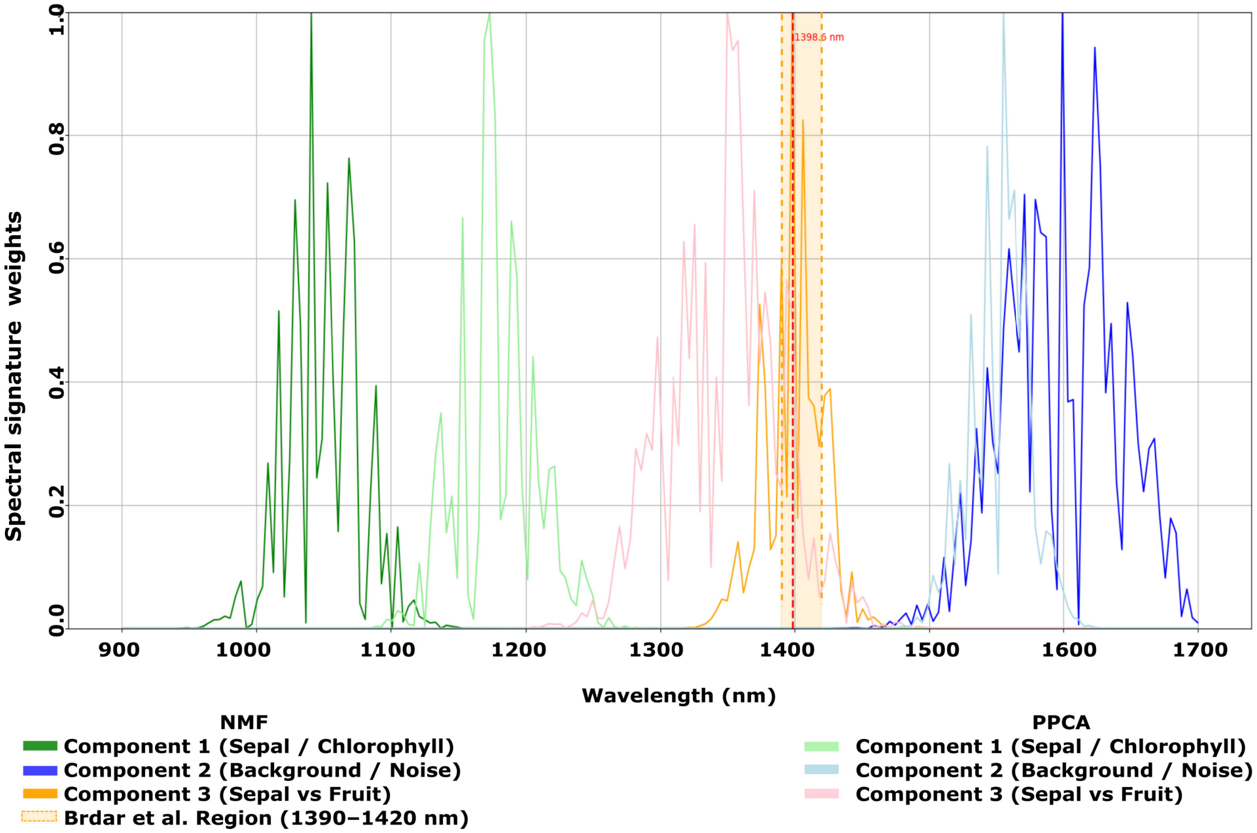

FIGURE 8. Normalized NMF and PPCA spectral encoded Dataset1 HSI of three principal components. A dominant peak in Component 3, obtained by NMF, is observed at 1398.6 nm, overlapping with the 1390–1420 nm region (highlighted) previously reported by [32] as significant for water-related differentiation.

susceptible to segmentation errors but retain the most critical information [124]. Even minor inaccuracies in these regions could lead to the loss of crucial morphological information, potentially affecting the robustness of susceptibility assessments. In this analysis, we evaluated and compared the dominant wavelength contributions represented in the top three principal components derived from NMF, which was identified as the best-performing method in our previously presented results table, followed by PPCA (see Table 3) and Table 7). These components were extracted from the raw hyperspectral data (Dataset1) to identify the most significant spectral features and compare them with the findings from [32].

The components were plotted across the full spectral range of 900–1700 nm, with each curve normalized to facilitate comparison (Figure 8). A clear peak was observed in Component 3, obtained by NMF, around 1398.6 nm, aligning well with the range identified by [32] as critical for water content discrimination in tomato tissues. This correspondence suggests that our data-driven NMF spectral encoding reinforces findings from the previous study and further validates the utility of this spectral window for distinguishing between tomato sepals and tomato fruit. Additionally, Component 1, by NMF, exhibited a prominent peak near 1050 nm, commonly associated with the estimation of chlorophyll content [125], [126] and indicative of sepal-related spectral features, while Component 2 by NMF showed a peak at 1592 nm, which is typically attributed to noise or background signals. Conversely, the PPCA components are slightly shifted compared to those derived from NMF. Component 1 showed a peak at 1180 nm, suggesting a broader spectral response of the sepal or overlapping biochemical signals. Component 2 peaked at 1550 nm, associated with water content and tissue scattering. Component 3 exhibited a peak at 1350 nm, reflecting structural differences between plant organs, such as tomato sepals and tomato fruit surfaces. These shifts are expected due to the probabilistic nature of PPCA, which blends signals across adjacent spectral bands, resulting in more distributed peaks than NMF. These findings, support the hypothesis that the 1398.6 nm region is particularly dominant in hyperspectral segmentation tasks, especially for distinguishing moisture-rich fruit from relatively drier sepals. The obtained results are consistent with those reported in [32], which identified the near-infrared (NIR) range between 1390–1420 nm as critical for separating water-rich and dry tissue zones.

Moreover, to evaluate the precision of the segmentation of the best-performing ConvMaskFPN model, initializing it with NMF coefficients, we calculated the mean and standard deviation for each wavelength (i.e., features) of

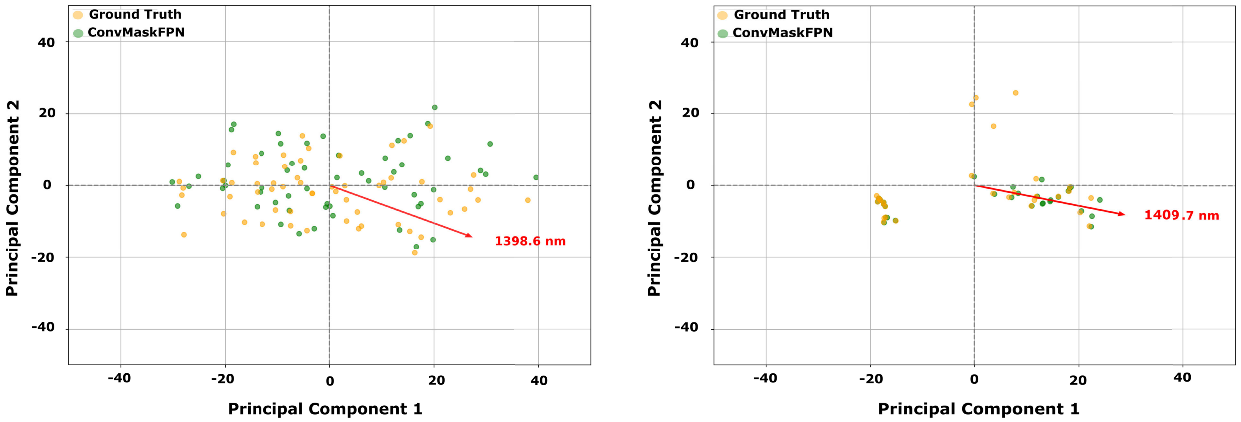

FIGURE 9. Dominant wavelengths contributing to the PCA projection formed from datasets of features calculated using ground truth masks and obtained from ConvMaskFPN model on the test sets on (a) Dataset 1 and (b) Dataset 2.

the original 196-dimensional hyperspectral response from both datasets per sepal by selecting those pixels per sepal denoted by the ground-truth mask and the estimated mask obtained by the selected model on test data. The Pearson correlation coefficients between vectorized values of features inferred from the ground-truth masks and masks generated by the ConvMaskFPN are: for Dataset 1 $r = 0.8100$, with a p-value of $p = 1.1940 \times 10^{-9}$ and for Dataset 2 $r = 0.7523$, with a p-value of $p = 6.2944 \times 10^{-9}$, which further confirms the high overall accuracy and precision of tomato sepal instance segmentation. Although we obtained quite a high correlation, precise pixel-level segmentation remains a challenge for this type of data, as each pixel carries a large amount of information, particularly when it is located at a key position, in this case, at the top of the tomato sepals. To further investigate the precision of the obtained sepals masks by ConvMaskFPN compared with the ground-truth masks, we conduct PCA for these two created datasets of features. Projecting the estimated feature vectors containing the means and standard deviations of each wavelength onto a two-dimensional subspace reveals, through the PCA loadings, the contribution of each feature to the creation of the PC component. In Figure 9, each arrow represents the highest contribution of a feature to the both principal components for both created datasets from the ground-truth and estimated masks. The most contributing features are similar. This region likely plays a significant role in guiding the model's decision-making process to predict the susceptibility of freshly harvested tomatoes and sepals to future fungal infections. Unveiling NMF components, along with obtaining results with ConvMaskFPN, demonstrates that this spectral region also provides valuable information for the sepals' segmentation task. In future work, we will investigate more deeply the causes of uncertainty in the detection of highly complex tomato sepal structures that could lead to improvements in model training or data preprocessing strategies. Although proposed spectral information encoding techniques are computationally efficient and easy to implement, [127] they have limitations. Further, advanced unsupervised deep learning-based techniques, such as autoencoders [128], [129] and variational autoencoders (VAEs) [130], [131] can improve the spectral information encoding step for hyperspectral images by learning data representations that are not limited to linear combinations of original features. This will enable the unveiling of more complex, nonlinear relationships in the data within the spectral dimension. Additionally, we intend to investigate self-supervised segmentation approaches, which offer promising potential in domains where generating high-quality labeled data is challenging, such as in hyperspectral imaging with domain-specific classes.

V. CONCLUSION

This research study demonstrates the effectiveness of tomato sepal instance segmentation within hyperspectral images by integrating well-established techniques PCA, PPCA, ICA, and NMF for encoding spectral information, with pre-trained deep learning models. Transfer learning proved to be a valuable approach, as models pre-trained on the COCO dataset and fine-tuned on the hyperspectral images with encoded spectral information achieved outstanding performances. Among the tested models, Mask R-CNN with the FPN backbone and NMF technique for encoding spectral information (NMFMaskFPN), showed the most stable and reliable performance, providing accurate segmentations with a low standard deviation of mAP. Additionally, initializing the weights of the added layer with pre-trained coefficients, obtained through spectral encoding with the NMF technique, led to a slight improvement in overall precision. This marks progress towards developing an end-to-end framework for encoding spectral information within hyperspectral imagery and further segmenting regions of interest within them. The variation in over- and under-prediction of sepals across different models and encoding techniques further emphasized the need for careful selection and optimization of models for specific tomato structures. The findings of this study show the possibilities for advanced practical applications of hyperspectral imaging in the fields of post-harvest supply chain management, including quality assessment, defect detection, and monitoring of the consequences of storage conditions on harvested tomatoes.

Further research could address in more detail the optimization of parameters for transfer learning and explore more advanced methods for spectral information encoding tailored to the unique characteristics of hyperspectral data. Moreover, such research might concentrate on refining and customizing the proposed model for specific tomato varieties or adapting it to address challenges posed by varying environmental conditions. The presented interdisciplinary research not only provides valuable insights into the current state of hyperspectral imaging for tomato analysis but also suggests future avenues of research and practical applications within the agricultural domain.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to Esther Van Echtelt for her valuable contributions and expertise in the early stage of this study. Her support in providing datasets has been greatly appreciated.

REFERENCES

[1] B. Farneti, "Tomato quality: From the field to the consumer: Interactions between genotype, cultivation and postharvest conditions," Ph.D. dissertation, Wageningen Food Biobased Res., Wageningen Univ. Res., 2014.

[2] J. Janse and H. Boerrigter, "Kroonschimmel bij tomaat: Consultancyonderzoek," Wageningen UR Glastuinbouw, Bleiswijk, The Netherlands, Tech. Rep. 13-002, 2007.

[3] E. J. Smid, L. Hendriks, H. A. M. Boerrigter, and L. G. M. Gorris, "Surface disinfection of tomatoes using the natural plant compound trans-cinnamaldehyde," *Postharvest Biol. Technol.*, vol. 9, no. 3, pp. 343–350, Dec. 1996.

[4] M. Liu and P. C. Ma, "Postharvest problems of vegetables and fruits in the tropics and subtropics," Asian Vegetable Res. Develop., Center, Shanhua, Tainan, Taiwan, Tech. Rep., 1983.

[5] E. E. Stinson, S. F. Osman, E. G. Heisler, J. Siciliano, and D. D. Bills, "Mycotoxin production in whole tomatoes, apples, oranges, and lemons," *J. Agricult. Food Chem.*, vol. 29, no. 4, pp. 790–792, Jul. 1981.

[6] J. Ukeh and N. Chiejina, "Preliminary investigations of the cause of postharvest fungal rot of Tomato," *IOSR J. Pharmacy Biol. Sci.*, vol. 4, no. 5, pp. 36–39, 2012.

[7] G. Agrios, *Plant Pathology*. Amsterdam, The Netherlands: Elsevier, 2005.

[8] F. A. A. Ibrahim, "Evaluation of antifungal activity of some plant extracts and their applicability in extending the shelf life of stored tomato fruits," *J. Food Process. Technol.*, vol. 5, no. 6, p. 340, 2014.

[9] Y. Zhang, R. De Stefano, M. Robine, E. Butelli, K. Bulling, L. Hill, M. Rejzek, C. Martin, and H.-j. Schoonbeek, "Different reactive oxygen species scavenging properties of flavonoids determine their abilities to extend the shelf life of tomato," *Plant Physiol.*, vol. 169, no. 3, pp. 1568–1583, 2015.

[10] K. Pobiega, J. L. Przybył, J. Żubernik, and M. Gniewosz, "Prolonging the shelf life of cherry tomatoes by pullulan coating with ethanol extract of propolis during refrigerated storage," *Food Bioprocess Technol.*, vol. 13, no. 8, pp. 1447–1461, Aug. 2020.

[11] S. K. S. Durai and M. D. Shamili, "Smart farming using machine learning and deep learning techniques," *Decis. Analytics J.*, vol. 3, Jun. 2022, Art. no. 100041.

[12] T. Ayoub Shaikh, T. Rasool, and F. Rasheed Lone, "Towards leveraging the role of machine learning and artificial intelligence in precision agriculture and smart farming," *Comput. Electron. Agricult.*, vol. 198, Jul. 2022, Art. no. 107119.

[13] S. Prabakaran, T. Dharun, H. Gowsik, A. Prakash, and K. Kirubakaran, "Smart farm management using machine learning," in *Proc. Int. Conf. IoT, Commun. Autom. Technol. (ICICAT)*, Nov. 2024, pp. 225–229.

[14] J. S. Prashanth, G. B. Krishna, A. V. K. Prasad, and P. R. Rao, "Smart farming revolution: A cutting-edge review of deep learning and IoT innovations in agriculture," in *Proc. Operations Res. Forum*, Mar. 2025, vol. 6, no. 1, pp. 1–39.

[15] L. Yan, M. Zhao, X. Wang, Y. Zhang, and J. Chen, "Object detection in hyperspectral images," *IEEE Signal Process. Lett.*, vol. 28, pp. 508–512, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1051200422003694

[16] S. Ding, H. Zhao, Y. Zhang, X. Xu, and R. Nie, "Extreme learning machine: Algorithm, theory and applications," *Artif. Intell. Rev.*, vol. 44, no. 1, pp. 103–115, Jun. 2015.

[17] M. Zhang, Z. Qin, X. Liu, and S. L. Ustin, "Detection of stress in tomatoes induced by late blight disease in California, USA, using hyperspectral remote sensing," *Int. J. Appl. Earth Observ. Geoinformation*, vol. 4, no. 4, pp. 295–310, Nov. 2003.

[18] A. A. Nielsen, "Kernel maximum autocorrelation factor and minimum noise fraction transformations," *IEEE Trans. Image Process.*, vol. 20, no. 3, pp. 612–624, Mar. 2011.

[19] A. Sachinopoulou, "Multidimensional visualization," VTT Tech. Res. Centre Finland, Espoo, Finland, Res. Notes VTT Research Notes 2114, 2001. [Online]. Available: https://publications.vtt.fi/pdf/tiedotteet/2001/T2114.pdf

[20] P. J. Martínez, R. M. Pérez, A. Plaza, P. L. Aguilar, M. C. Cantero, and J. Plaza, "End member extraction algorithms from hyperspectral images," *Ann. Geophysics*, vol. 49, no. 1, pp. 108–119, Dec. 2009. [Online]. Available: https://api.semanticscholar.org/CorpusID

[21] F. A. Kruse, A. B. Lefkoff, and J. B. Dietz, "Expert system-based mineral mapping in northern death valley, California/nevada, using the airborne Visible/Infrared imaging spectrometer (AVIRIS)," *Remote Sens. Environ.*, vol. 44, nos. 2–3, pp. 309–336, May 1993.

[22] Q. Lü and M. Tang, "Detection of hidden bruise on kiwi fruit using hyperspectral imaging and parallelepiped classification," *Proc. Environ. Sci.*, vol. 12, pp. 1172–1179, Jan. 2012.

[23] J. A. Richards and X. Jia, "The interpretation of digital image data," in *Proc. Remote Sens. Digit. Image Analysis, Introduction*, Jan. 1999, pp. 75–88.

[24] Y. Jiang, C. Li, and F. Takeda, "Nondestructive detection and quantification of blueberry bruising using near-infrared (NIR) hyperspectral reflectance imaging," *Sci. Rep.*, vol. 6, no. 1, pp. 1–14, Oct. 2016.

[25] W. S. Noble, "What is a support vector machine?" *Nature Biotechnol.*, vol. 24, no. 12, pp. 1565–1567, Dec. 2006.

[26] S. Fan, C. Li, W. Huang, and L. Chen, "Data fusion of two hyperspectral imaging systems with complementary spectral sensing ranges for blueberry bruising detection," *Sensors*, vol. 18, no. 12, p. 4463, Dec. 2018.

[27] H. Ma, K. Zhao, X. Jin, J. Ji, Z. Qiu, and S. Gao, "Spectral difference analysis and identification of different maturity blueberry fruit based on hyperspectral imaging using spectral index," *Int. J. Agricult. Biol. Eng.*, vol. 12, no. 3, pp. 134–140, 2019.

[28] Y. Huang, D. Wang, Y. Liu, H. Zhou, and Y. Sun, "Measurement of early disease blueberries based on Vis/NIR hyperspectral imaging system," *Sensors*, vol. 20, no. 20, p. 5783, Oct. 2020.

[29] S. Qiao, Q. Wang, J. Zhang, and Z. Pei, "Detection and classification of early decay on blueberry based on improved deep residual 3D convolutional neural network in hyperspectral images," *Scientific Program.*, vol. 2020, pp. 1–12, May 2020.

[30] C. Zhou, Z. Li, D. Wang, S. Xue, T. Zhu, and C. Ni, "SSNet: Exploiting spatial information for tobacco stem impurity detection with hyperspectral imaging," *IEEE Access*, vol. 12, pp. 55134–55145, 2024.

[31] J. Li, F. Xu, S. Song, Q. Ji, and J. Liu, "Hyperspectral RGB imaging combined with deep learning for maize seed variety identification," *IEEE Access*, vol. 12, pp. 184477–184486, 2024.

[32] S. Brdar, M. Panić, E. Hogeveen-Van Echtelt, M. Mensink, Ž. Grbović, E. Woltering, and A. Chauhan, "Predicting sensitivity of recently harvested tomatoes and tomato sepals to future fungal infections," *Sci. Rep.*, vol. 11, no. 1, p. 23109, Nov. 2021.

[33] M. E. Paoletti, J. M. Haut, J. Plaza, and A. Plaza, "Deep learning classifiers for hyperspectral imaging: A review," *ISPRS J. Photogramm. Remote Sens.*, vol. 158, pp. 279–317, Dec. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271619302187

[34] X. He, C. Tang, X. Liu, W. Zhang, K. Sun, and J. Xu, "Object detection in hyperspectral image via unified spectral–spatial feature aggregation," 2023, *arXiv:2306.08370*.

[35] Q. Wang, F. Zhang, and X. Li, "Hyperspectral band selection via optimal neighborhood reconstruction," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8465–8476, Dec. 2020.

[36] Sneha and A. Kaul, "Hyperspectral imaging and target detection algorithms: A review," *Multimedia Tools Appl.*, vol. 81, no. 30, pp. 44141–44206, Dec. 2022.

[37] H. Chen, F. Miao, Y. Chen, Y. Xiong, and T. Chen, "A hyperspectral image classification method using multifeature vectors and optimized KELM," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2781–2795, 2021.

[38] C. Xie, Y. Shao, X. Li, and Y. He, "Detection of early blight and late blight diseases on tomato leaves using hyperspectral imaging," *Sci. Rep.*, vol. 5, no. 1, pp. 1–11, Nov. 2015.

[39] S. F. C. Soares, A. A. Gomes, M. C. U. D. Araãjo, A. R. G. Filho, and R. K. H. Galvão, "The successive projections algorithm," *TrAC Trends Anal. Chem.*, vol. 42, pp. 84–98, Oct. 2012.

[40] K. Makantasis, K. Karantzalos, A. Doulamis, and N. Doulamis, "Deep supervised learning for hyperspectral data classification through convolutional neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2015, pp. 4959–4962.

[41] L. Yan, M. Yamaguchi, N. Noro, Y. Takara, and F. Ando, "A novel two-stage deep learning-based small-object detection using hyperspectral images," *Opt. Rev.*, vol. 26, no. 6, pp. 597–606, Dec. 2019.

[42] S. Ruder, "An overview of gradient descent optimization algorithms," 2016, *arXiv:1609.04747*.

[43] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*. Amsterdam, The Netherlands: Elsevier, 2006.

[44] A. Kallepalli, A. Kumar, and K. Khoshelham, "Entropy based determination of optimal principal components of airborne prism experiment (APEX) imaging spectrometer data for improved land cover classification," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vols. XL–8, pp. 781–786, Nov. 2014. [Online]. Available: https://isprs-archives.copernicus.org/articles/XL-8/781/2014/

[45] C. Sima and E. R. Dougherty, "The peaking phenomenon in the presence of feature-selection," *Pattern Recognit. Lett.*, vol. 29, no. 11, pp. 1667–1674, Aug. 2008.

[46] A. Signoroni, M. Savardi, A. Baronio, and S. Benini, "Deep learning meets hyperspectral image analysis: A multidisciplinary review," *J. Imag.*, vol. 5, no. 5, p. 52, May 2019.

[47] W. Li, G. Wu, F. Zhang, and Q. Du, "Hyperspectral image classification using deep pixel-pair features," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 844–853, Feb. 2017.

[48] L. Ran, Y. Zhang, W. Wei, and Q. Zhang, "A hyperspectral image classification framework with spatial pixel pair features," *Sensors*, vol. 17, no. 10, p. 2421, Oct. 2017.

[49] W. Sun and Q. Du, "Hyperspectral band selection: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 118–139, Jun. 2019.

[50] A. A. Adedeji, N. Ekramirad, A. Rady, A. Hamidisepehr, K. D. Donohue, R. T. Villanueva, C. A. Parrish, and M. Li, "Non-destructive technologies for detecting insect infestation in fruits and vegetables under postharvest conditions: A critical review," *Foods*, vol. 9, no. 7, p. 927, Jul. 2020.

[51] A. Singh, G. Vaidya, V. Jagota, D. A. Darko, R. K. Agarwal, S. Debnath, and E. Potrich, "Recent advancement in postharvest loss mitigation and quality management of fruits and vegetables using machine learning frameworks," *J. Food Qual.*, vol. 2022, pp. 1–9, Jun. 2022.

[52] G. Dyck, E. Hawley, K. Hildebrand, and J. Paliwal, "Digital twins: A novel traceability concept for post-harvest handling," *Smart Agricult. Technol.*, vol. 3, Feb. 2023, Art. no. 100079.

[53] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[54] A. Khamparia and K. M. Singh, "A systematic review on deep learning architectures and applications," *Expert Syst.*, vol. 36, no. 3, p. 12400, Jun. 2019.

[55] A. Shrestha and A. Mahmood, "Review of deep learning algorithms and architectures," *IEEE Access*, vol. 7, pp. 53040–53065, 2019.

[56] A. Mathew, P. Amudha, and S. Sivakumari, "Deep learning techniques: An overview," in *Proc. Adv. Mach. Learn. Technol. Appl.*, May 2020, pp. 599–608.

[57] M. Gheisari, F. Ebrahimzadeh, M. Rahimi, M. Moazzamigodarzi, Y. Liu, P. K. Dutta Pramanik, M. A. Heravi, A. Mehbodniya, M. Ghaderzadeh, M. R. Feylizadeh, and S. Kosari, "Deep learning: Applications, architectures, models, tools, and frameworks: A comprehensive survey," *CAAI Trans. Intell. Technol.*, vol. 8, no. 3, pp. 581–606, Sep. 2023.

[58] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[59] P. O. Pinheiro, T.-Y. Lin, R. Collobert, and P. Dollár, "Learning to refine object segments," in *Proc. 14th Eur. Conf. Comput. Vision ECCV*, Amsterdam, The Netherlands, Jan. 2016, pp. 75–91.

[60] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[61] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, "Deep learning for generic object detection: A survey," 2018, *arXiv:1809.02165*.

[62] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[63] S. Zagoruyko, A. Lerer, T.-Y. Lin, P. O. Pinheiro, S. Gross, S. Chintala, and P. Dollár, "A MultiPath network for object detection," 2016, *arXiv:1604.02135*.

[64] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, Dec. 2015, pp. 91–99.

[65] K. He, G. Gkioxari, P. Dollr, and R. Girshick, "Mask R-CNN," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 386–397, Feb. 2020.

[66] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[67] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8759–8768.

[68] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, "Hybrid task cascade for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4969–4978.

[69] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1971–1980.

[70] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "YOLACT: Real-time instance segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9156–9165.

[71] X. Chen, R. Girshick, K. He, and P. Dollar, "TensorMask: A foundation for dense object segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2061–2069.

[72] J. Lazarow, W. Xu, and Z. Tu, "Instance segmentation with mask-supervised polygonal boundary transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4372–4381.

[73] W. Chen, X. Du, F. Yang, L. Beyer, X. Zhai, T.-Y. Lin, H. Chen, J. Li, X. Song, Z. Wang, and D. Zhou, "A simple single-scale vision transformer for object detection and instance segmentation," in *Proc. Eur. Conf. Comput. Vis.*, Jan. 2022, pp. 711–727.

[74] S. K. Zhou, H. Greenspan, C. Davatzikos, J. S. Duncan, B. Van Ginneken, A. Madabhushi, J. L. Prince, D. Rueckert, and R. M. Summers, "A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises," *Proc. IEEE*, vol. 109, no. 5, pp. 820–838, May 2021.

[75] S. Mozaffari, O. Y. Al-Jarrah, M. Dianati, P. Jennings, and A. Mouzakitis, "Deep learning-based vehicle behavior prediction for autonomous driving applications: A review," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 1, pp. 33–47, Jan. 2022.

[76] S. Coulibaly, B. Kamsu-Foguem, D. Kamissoko, and D. Traore, "Deep learning for precision agriculture: A bibliometric analysis," *Intell. Syst. Appl.*, vol. 16, Nov. 2022, Art. no. 200102.

[77] Y. Chen, X. Zhao, and X. Jia, "Spectral–Spatial classification of hyperspectral data based on deep belief network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2381–2392, Jun. 2015.

[78] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6232–6251, Oct. 2016.

[79] X. Cao, J. Yao, Z. Xu, and D. Meng, "Hyperspectral image classification with convolutional neural network and active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4604–4616, Jul. 2020.

[80] Y. Luo, J. Zou, C. Yao, X. Zhao, T. Li, and G. Bai, "HSI-CNN: A novel convolution neural network for hyperspectral image," in *Proc. Int. Conf. Audio, Lang. Image Process. (ICALIP)*, Jul. 2018, pp. 464–469.

[81] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6690–6709, Sep. 2019.

[82] Z. Zhong, J. Li, L. Ma, H. Jiang, and H. Zhao, "Deep residual networks for hyperspectral image classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2017, pp. 1824–1827.

[83] C. Debes, A. Merentitis, R. Heremans, J. Hahn, N. Frangiadakis, T. van Kasteren, W. Liao, R. Bellens, A. Pižurica, S. Gautama, W. Philips, S. Prasad, Q. Du, and F. Pacifici, "Hyperspectral and LiDAR data fusion: Outcome of the 2013 GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2405–2418, Jun. 2014.

[84] X. Cao, F. Zhou, L. Xu, D. Meng, Z. Xu, and J. Paisley, "Hyperspectral image classification with Markov random fields and a convolutional neural network," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2354–2367, May 2018.

[85] Y. Chen, L. Lin, and L. Chen, "Attention-based spectral–spatial classification of hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 844–853, Feb. 2017.

[86] D. S. Joseph, P. M. Pawar, and R. Pramanik, "Intelligent plant disease diagnosis using convolutional neural network: A review," *Multimedia Tools Appl.*, vol. 82, no. 14, pp. 21415–21481, Jun. 2023.

[87] A. N. N. Azmi, S. K. Bejo, M. Jahari, and I. J. Yule, "Early detection of plant disease infection using hyperspectral data and machine learning," in *Proc. IoT AI Agricult.*, Jan. 2023, pp. 423–446.

[88] I. T. Jolliffe, *Principal Component Analysis for Special Types of Data*. Cham, Switzerland: Springer, 2002.

[89] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *J. Educ. Psychol.*, vol. 24, no. 6, pp. 417–441, Sep. 1933.

[90] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *J. Roy. Stat. Soc. B, Stat. Methodology*, vol. 61, no. 3, pp. 611–622, 1999.

[91] A. T. Basilevsky, *Statistical Factor Analysis and Related Methods: Theory and Applications*. Hoboken, NJ, USA: Wiley, 2009.

[92] D. Barber, *Bayesian Reasoning and Machine Learning*. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[93] C. M. Bishop and N. M. Nasrabadi, *Pattern Recognition and Machine Learning*, vol. 4. Cham, Switzerland: Springer, 2006.

[94] V. K. Munipalle, U. Rani Nelakuditi, and R. R. Nidamanuri, "Impact of dimensionality reduction techniques on classification of hyperspectral images," in *Proc. 3rd Int. Conf. Intell. Technol. (CONIT)*, Jun. 2023, pp. 1–6.

[95] M. Lennon, G. Mercier, M. C. Mouchot, and L. Hubert-Moy, "Independent component analysis as a tool for the dimensionality reduction and the representation of hyperspectral images," in *Proc. Scanning Present Resolving Future. IEEE Int. Geosci. Remote Sens. Symp.*, vol. 6, Jul. 2001, pp. 2893–2895.

[96] J. Wang and C.-I. Chang, "Independent component analysis-based dimensionality reduction with applications in hyperspectral image analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 6, pp. 1586–1600, Jun. 2006.

[97] S. B. Pena, M. M. Abreu, and M. R. Magalhães, "Planning landscape with water Infiltration. Empirical model to assess maximum infiltration areas in Mediterranean landscapes," *Water Resour. Manage.*, vol. 30, no. 7, pp. 2343–2360, May 2016.

[98] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, Jan. 2014, pp. 740–755.

[99] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[100] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, Jan. 2021.

[101] L. Torrey and J. Shavlik, "Transfer learning," in *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*. Hershey, PA, USA: IGI global, 2010, pp. 242–264.

[102] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[103] Z. U. Rehman, M. A. Khan, F. Ahmed, R. Damaševičius, S. R. Naqvi, W. Nisar, and K. Javed, "Recognizing apple leaf diseases using a novel parallel real-time processing framework based on MASK RCNN and transfer learning: An application for smart agriculture," *IET Image Process.*, vol. 15, no. 10, pp. 2157–2168, Aug. 2021.

[104] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," 2017, *arXiv:1703.06211*.

[105] H.-J. He, Y. Chen, G. Li, Y. Wang, X. Ou, and J. Guo, "Hyperspectral imaging combined with chemometrics for rapid detection of talcum powder adulterated in wheat flour," *Food Control*, vol. 144, Feb. 2023, Art. no. 109378.

[106] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: A database and Web-based tool for image annotation," *Int. J. Comput. Vis.*, vol. 77, nos. 1–3, pp. 157–173, May 2008.

[107] T. Boggs. (2014). *Spectral Python (spy)*. [Online]. Available: http://www.spectralpython.net

[108] S. van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, and T. Yu, "Scikit-image: Image processing in Python," *PeerJ*, vol. 2, p. e453, Jun. 2014.

[109] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. J. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Jan. 2012.

[110] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick. (2019). *Detectron2*. [Online]. Available: https://github.com/facebookresearch/detectron2

[111] R. Padilla, W. L. Passos, T. L. B. Dias, S. L. Netto, and E. A. B. da Silva, "A comparative analysis of object detection metrics with a companion open-source toolkit," *Electronics*, vol. 10, no. 3, p. 279, Jan. 2021.

[112] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2016, pp. 1–11.

[113] A. Banerjee, U. Chitnis, S. Jadhav, J. Bhawalkar, and S. Chaudhury, "Hypothesis testing, type i and type II errors," *Ind. Psychiatry J.*, vol. 18, no. 2, p. 127, 2009.

[114] Y. Benjamini, "Discovering the false discovery rate," *J. Roy. Stat. Soc. B, Stat. Methodology*, vol. 72, no. 4, pp. 405–416, Sep. 2010.

[115] B. Galindo-Prieto and F. Westad, "Classification in hyperspectral images by independent component analysis, segmented cross-validation and uncertainty estimates," *J. Spectral Imag.*, vol. 7, pp. 1–21, Feb. 2018.

[116] X.-R. Feng, H.-C. Li, R. Wang, Q. Du, X. Jia, and A. Plaza, "Hyperspectral unmixing based on nonnegative matrix factorization: A comprehensive review," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 4414–4436, 2022.

[117] H. Wei and J. P. Smith, "Machine learning for deconvolution and segmentation of hyperspectral imaging data from biopharmaceutical resins," *Mol. Pharmaceutics*, vol. 21, no. 11, pp. 5565–5576, Nov. 2024.

[118] N. Shirish Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. Tak Peter Tang, "On large-batch training for deep learning: Generalization gap and sharp minima," 2016, *arXiv:1609.04836*.

[119] S. L. Smith, P.-J. Kindermans, C. Ying, and Q. V. Le, "Don't decay the learning rate, increase the batch size," *arXiv:1711.00489*, 2017.

[120] Z. Liu, "Super convergence cosine annealing with warm-up learning rate," in *Proc. CAIBDA ; 2nd Int. Conf. Artif. Intell., Big Data Algorithms*, Jun. 2022, pp. 1–7.

[121] Z. Xu, A. M. Dai, J. Kemp, and L. Metz, "Learning an adaptive learning rate schedule," 2019, *arXiv:1909.09712*.

[122] A. Defazio, A. Cutkosky, H. Mehta, and K. Mishchenko, "Optimal linear decay learning rate schedules and further refinements," 2023, *arXiv:2310.07831*.

[123] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.

[124] V. Filipovic, Ž. Grbović, A. Chauhan, H. de Villiers, M. Panić, and S. Brdar, "Influence of spatial region selection on infection sensitivity prediction of tomato calyx using hyperspectral imaging," in *Proc. 14th Int. Conf. Inf. Soc. Technol.*, 2024, pp. 1–21.

[125] Z. Wang, X. Tan, Y. Ma, T. Liu, L. He, F. Yang, C. Shu, L. Li, H. Fu, B. Li, Y. Sun, Z. Yang, Z. Chen, and J. Ma, "Combining canopy spectral reflectance and RGB images to estimate leaf chlorophyll content and grain yield in rice," *Comput. Electron. Agricult.*, vol. 221, Jun. 2024, Art. no. 108975.

[126] P. Huang, P. Yang, L. Yang, F. Xiao, Y. Feng, and Y. Wang, "Non-destructive detection and visualization of chlorophyll content in cherry tomatoes based on hyperspectral technology and machine learning," *Agriculture*, vol. 14, no. 12, p. 2247, Dec. 2024.

[127] B. Ghojogh, A. Ghodsi, F. Karray, and M. Crowley, "Factor analysis, probabilistic principal component analysis, variational inference, and variational autoencoder: Tutorial and survey," 2021, *arXiv:2101.00734*.

[128] P. Baldi, "Autoencoders, unsupervised learning and deep architectures," in *Proc. ICML Workshop Unsupervised Transf. Learn.*, Jul. 2011, pp. 37–50.

[129] G. Jaiswal, R. Rani, H. Mangotra, and A. Sharma, "Integration of hyperspectral imaging and autoencoders: Benefits, applications, hyperparameter tunning and challenges," *Comput. Sci. Rev.*, vol. 50, Nov. 2023, Art. no. 100584.

[130] D. P. Kingma and M. Welling, "An introduction to variational autoencoders," *Found. Trends Mach. Learn.*, vol. 12, no. 4, pp. 307–392, 2019.

[131] C. Loughlin, D. Manolakis, M. Pieper, V. Ingle, R. Bostick, and A. Weisner, "Spectral variability modeling with variational autoencoders for hyperspectral target analysis," *Proc. SPIE*, vol. 12519, pp. 214–226, May 2023.

ŽELJANA GRBOVIĆ received the M.Sc. degree in electrical engineering from the Faculty of Technical Sciences, University of Novi Sad, Serbia, in 2017. She is currently pursuing the Ph.D. degree in electrical and computer engineering with the Faculty of Technical Sciences. Since 2017, she has been a Researcher with the BioSense Institute. Her research interests include processing various image modalities (thermal, hyperspectral, and RGB) involving computer vision techniques, such as 3-D reconstruction with application in agricultural practices. She was a member of the team that placed as a finalist in two OpenCV AI Spatial competitions and a recipient of the Best Female Engineer of the Year 2024 Award from the Chamber of Commerce of Serbia and Siemens.

MARKO PANIĆ received the dual Ph.D. degree from the University of Novi Sad and Ghent University, in 2020. He is a Senior Research Associate with the BioSense Institute, University of Novi Sad, Serbia. He has been actively involved in international collaborations, participating in several HORIZON2020 EU-funded projects. His research focuses on compressed sensing theory and multimodal imaging sensors, with applications in biology, agriculture, environmental sciences, and health. As a member of the BioSense Team, he has competed the Syngenta Crop Challenge (three times), securing first and third place, and was a finalist in two OpenCV challenges.

VLADAN FILIPOVIĆ received the M.Sc. degree from the Faculty of Technical Sciences, University of Novi Sad, in 2021, where he is currently pursuing the Ph.D. degree in electrical and computer engineering. He is a Research Assistant with the BioSense Institute. His research interests include the application of artificial intelligence, image processing, and computer vision in the fields of precision agriculture and postharvest technology. Notable areas of his research include hyperspectral image analysis for early detection of fungal infection in fruit and varietal classification of grain seeds. Besides these, his research includes RGB-depth image processing for predicting crop yields and visual robotic perception for various tasks, such as guiding remote-controlled ground vehicles (RGVs) and weed spraying in blueberry orchards. As a member of the BioSense Team, he has participated in two OpenCV challenges, placing as a finalist both times.

SANJA BRDAR received the Ph.D. degree in electrical and computer engineering from the Faculty of Technical Sciences, University of Novi Sad, Serbia, in 2016. She is a Research Associate Professor with the BioSense Institute, University of Novi Sad. Her research interests include artificial intelligence and bioinformatics. She works on ensemble methods, data fusion, clustering, and predictive modeling with applications in biology, agriculture, and environmental sciences. On worldwide data science challenges with the team of researchers, she placed third in Nokia Mobile Data Challenge, in 2012; a finalist of Orange, France, Data for Development Challenge, in 2013; a finalist/winner/third place of Syngenta Crop Challenge, in 2016, 2017, and 2019; and a winner of Copernicus Masters–Space for Smart Mobility Challenge, in 2021.

HENDRIK DE VILLIERS received the Ph.D. degree in electronic engineering from the University of Stellenbosch, in 2014. In 2016, he joined Wageningen Food and Biobased Research, as a Computer Vision Researcher, works on topics including the quality grading of produce and plant disease detection using RGB and hyperspectral imaging and from point clouds. He is also involved in biodiversity monitoring projects. His current research interests include deep neural networks for self-supervised learning, few-shot learning, and anomaly detection.

MANON MENSINK received the B.S. degree in botanical laboratory technology in Wageningen, The Netherlands, in 1988. She is currently a Researcher with the Department of Postharvest Technology, Wageningen Food and Biobased Research. Her expertise is design of experiments (DoE) with regards to measurement of effect of treatments to maintain quality of fresh products.

ANEESH CHAUHAN received the M.Sc. degree in autonomous systems from the University of Exeter, U.K., in 2004, and the Ph.D. degree in informatics and robotics from the University of Aveiro, Portugal, in 2014. He has held postdoctoral positions with the University of Aveiro, on embodied robotics, and later with the Aerial Robotics Laboratory, UPM, Spain. Since 2017, he has been a Senior Scientist in computer vision and robotics with Wageningen Food and Biobased Research. His research interests include computer vision, machine (deep) learning, AI-enhanced robotics, human–robot interaction, and natural language processing. In his current role, he is applying this knowledge to the challenges in agri-food domains, such as tackling labor shortage, next-generation agri-food industry, autonomous robotics, personalized nutrition, human behavior understanding, and biodiversity monitoring.