Computers and Electronics in Agriculture 224 (2024) 109219

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

Varroa destructor detection on honey bees using hyperspectral imagery

Zina-Sabrina Duma [a,*], Tomas Zemcik [b], Simon Bilik [a,b], Tuomas Sihvonen [a], Peter Honec [b], Satu-Pia Reinikainen [a], Karel Horak [b]

[a] LUT University, Yliopistonkatu 34, Lappeenranta 53850, Finland
[b] Brno University of Technology, Faculty of Electrical Engineering and Communication, Technická 3058/10, Brno 61600, Czech Republic

ARTICLE INFO

Dataset link: http://dx.doi.org/10.34740/KAGGLE/DSV/7845514, https://www.kaggle.com/dsv/7845514

Keywords:
Hyperspectral imagery (HSI)
Varroa destructor
Beehive monitoring
Wavelength selection
Kernel partial least-squares

ABSTRACT

Hyperspectral (HS) imagery in agriculture is becoming increasingly common. These images have the advantage of higher spectral resolution. Advanced spectral processing techniques are required to unlock the information potential in these HS images. The present paper introduces a method rooted in multivariate statistics designed to detect parasitic *Varroa destructor* mites on the body of western honey bee *Apis mellifera*, enabling easier and continuous monitoring of the bee hives. The present paper is the first to utilize hyperspectral imagery for the task, previous studies existing only for multispectral imagery. The methodology explores unsupervised (K-means++) and recently developed supervised (Kernel Flows-Partial Least-Squares, KF-PLS) methods for parasitic identification. Additionally, in light of the emergence of custom-band multispectral cameras, the present research outlines a strategy for identifying the specific wavelengths necessary for effective bee-mite separation, suitable for implementation in a custom-band camera. Illustrated with a real-case dataset, our findings demonstrate that as few as four spectral bands are sufficient for accurate parasite identification.

1. Introduction

Environmental issues, alongside the use of pesticides and the presence of parasites, pose significant threats to bee populations worldwide. Among these threats, the *Varroa Destructor* (Varroa) mite is particularly notorious for its role in most instances of Colony Collapse Disorder (CCD), as highlighted in recent studies (Flores et al., 2021; Eliash and Mikheyev, 2020).

Traditionally, detecting Varroa mites within beehives has relied on manual, non-automated methods such as sugar shake tests, brood examinations, and debris analysis (Jack and Ellis, 2021; Roth et al., 2020). The recent advancements have introduced computer vision techniques (CV) utilizing the automated analysis of the bee debris on a monitoring plate in König (2020), or directly on the bee's body using conventional CV techniques in Bjerge et al. (2019), convolutional neural network (CNN) classifiers in Picek et al. (2022) or the deep object detectors in Bilik et al. (2021a), Liu et al. (2023). An extensive overview of the computer vision techniques used for the Varroa mite monitoring and bee colony health monitoring, in general, is shown, for example, in Bilik et al. (2024c). Besides the CV techniques, analysis based on the sensor and sound data is used, as presented, e.g. in Hall et al. (2023), Mekha et al. (2022). Nevertheless, reliable detection of the Varroa mite in the visible spectrum is challenging as it appears similar to the bee's body or the surroundings.

Hyperspectral (HS) imagery for agriculture monitoring is becoming readily available (Lu et al., 2020). There is an increased demand for precise performant HS imaging processing techniques (Khan et al., 2022). For monitoring bees and insects, HS imagery has been utilized previously in very few cases. More studies are found with multispectral data. The authors of Bjerge et al. (2019) used the VideometerLab4 instrument sensing in 19 wavelengths in the range of 375–970 nm to measure samples of the bees and Varroa mites followed by Linear Discriminant Analysis (LDA) on the spectral data to design an optimal illumination for their bee monitoring device; and the authors of Månefjord et al. (2022) used the multispectral data for insect monitoring, but not for Varroa mite detection.

In this study, we introduce a novel approach for extracting spectral data from HS images and leveraging this data to calibrate spectral signatures for identifying clusters of parasites. We also propose methods for selecting wavelengths that discriminate between clusters. The objective of this research is to explore the potential of HS imagery in addressing significant questions related to bee health and parasite detection:

- *Is it possible to use hyperspectral imagery to identify Varroa mites on bees?*
- *What procedures are necessary to extract discriminative information between bees and Varroa mites?*

* Corresponding author.
E-mail address: Zina-Sabrina.Duma@lut.fi (Z.-S. Duma).

https://doi.org/10.1016/j.compag.2024.109219
Received 21 March 2024; Received in revised form 27 June 2024; Accepted 1 July 2024
Available online 15 July 2024

- *How many and which specific wavelengths are crucial for distinguishing between bees and Varroa mites?*
- *Can statistical-based methods yield reliable results with independent data sets?*

Our methodology encompasses a process of spectral reconstruction aimed at enhancing the differentiation between bees and Varroa mites, which is crucial for supporting classification or clustering algorithms. This approach accentuates the contrasts between bee-mite characteristics while minimizing the variations caused by background elements, shadows, and pixel noise. Pixel noise, in particular, is often a byproduct of the line-scanning technique employed by hyperspectral cameras (Bjorgan and Randeberg, 2015). The process involves utilizing Principal Component Analysis (PCA) (Rodarmel and Shan, 2002) to decompose the HS image, followed by selecting only those principal components that demonstrate a strong absolute correlation with a bee-mite discrimination variable for the image reconstruction. K-means++ method (Hämäläinen et al., 2020) is applied to cluster the reconstructed HS images. Cluster centers are being applied to new images for effective discrimination.

Kernel Flows-Partial Least-Squares (KF-PLS) (Duma et al., 2023b) was utilized in the case of supervised clustering. It has the property to maintain the qualities of multivariate statistical methods, such as the reduced size of the calibration dataset, while extending applicability to non-linear relationships. The spectra of bees and parasites are not expected to be distinguished in a fully linear method; thus, there is a need for spectral projection in a Reproducing Kernel Hilbert Space, where they become linearly separable.

For the selection of spectral bands essential to discrimination between bees and Varroa mites, the mathematical methods utilized are two partial least-squares (PLS) (De Jong, 1993) based methods: a modified version of the Covariance Procedure (Reinikainen and Höskuldsson, 2003), mentioned throughout the paper as the COVPROC method, and the explained variance by wavelength (referred to as the R^2 method), derived from forward interval partial least-squares (FiPLS) (Yun et al., 2019).

The methodologies outlined in this paper were evaluated using a two-part proprietary hyperspectral image dataset. The methods utilized are all rooted in multivariate statistics, which has the advantage of requiring a very small data base for model calibration when compared to deep learning methods. This is due to the higher complexity and parameter tuning requirements of deep learning models, which need large datasets to avoid overfitting and ensure accurate predictions (Rajula et al., 2020). This dataset is made accessible alongside this article.

The novelty of the study lies in several key contributions: (*i*) introducing the use of HS imagery for detecting Varroa mites, a novel application in the field; (*ii*) developing a procedure for reconstructing spectral data that effectively excludes background details, shadows, and noise, ensuring clearer differentiation between subjects of interest; (*iii*) offering a new methodology for identifying discriminating spectral bands, which enhances understanding of the specific wavelengths vital for distinguishing between bees and Varroa mites; (*iv*) proposing a technique for creating spectral profile-based classifiers, which could improve how spectral data is used for classification purposes; (*v*) demonstrating the feasibility of using minimal data for training, addressing one of the common challenges in machine learning applications by reducing the requirement for extensive training datasets.

2. Materials and methods

This section presents the origin of bees and Varroa mites samples together with their measurement arrangement (Section 2.1), followed by the instrumentation, measurement setup and image data (Section 2.2), and summarized with the mathematical procedures and algorithms used for processing of the collected data (Section 2.3).

Table 1
Basic camera and image parameters of Specim IQ (Behmann et al., 2018; Ikäheimo and Jussila, 2018; Zemčík and Horak, 2023).

Specim IQ image parameters	
Spatial resolution	512 px × 512 px
Field of view	31° × 31°
F/number	1/1.7
Min. focus distance	150 mm
Scanning principle	push-broom
Spectral range	400 nm–1000 nm
Spectral resolution	7 nm
Spectral bands	204
Image format	ENVI compatible

2.1. Collected samples of bees and varroa mites

Two sets of samples were collected from different areas and at different times for the study. The first samples were taken from Těšínky (49.9566236N, 15.1436481E; CZ) in November 2021. The second was collected from Kroměříž (49.3005392N, 17.3797958E; CZ) in June 2022. These locations are situated approximately 180 km from one another. In both cases, the detritus containing dead mites and bees was collected from the bottom of the hives. In addition, detritus from the Těšínky locality was collected after a regular autumn fumigation with Amitrazum 125 mg/ml.

The first set of samples, illustrated in Fig. 1(a), 1(b), 1(d) and 1(e) contains samples arranged on the Petri dishes organized as follows:

- Bees mixed up with detritus from locality Kroměříž (Fig. 1(a)).
- Separated bees, detritus and Varroa mites from locality Kroměříž (Fig. 1(b)).
- Fresh bees from locality Kroměříž, marked as *K*, and eight months old bees from locality Těšínky, marked as *R* (Fig. 1(d)).
- Eight months old Varroa mites from locality Těšínky, marked as *O* and fresh Varroa mites from locality Kroměříž marked as *K* (Fig. 1(e)).

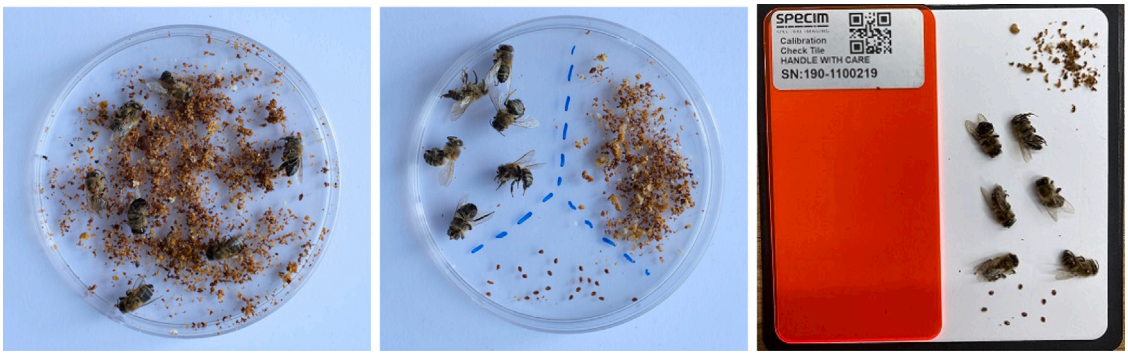
To suppress the effect of the polystyrene Petri dish used in HS images of the first set, a second set was prepared out of the Kroměříž samples placed directly on a white reference panel supplied with the used camera, which should have constant spectral properties across the camera's spectral range. Images of the detritus, bees, Varroa mites and bees with placed Varroa mites were taken in a dense and clustered arrangement as shown in the right column of Fig. 1, with the dense arrangement shown in Fig. 1(c) and clustered arrangement shown in Fig. 1(f).

In addition to the mites and the bees, the detritus consists of any kind of waste material from the nest, such as wax, pollen, or sugar, which is also visible in Fig. 1(a), 1(b), 1(c) and 1(f). In total, our data contain 40 bees and 72 Varroa mites. We also made samples from our dataset publicly available in Bilik et al. (2024b).

2.2. Hyperspectral imagery

The hyperspectral images were taken on a Specim IQ (Behmann et al., 2018) portable hyperspectral camera that allows a simple measurement setup and fast acquisition. Camera parameters relevant to the data format are included in Table 1.

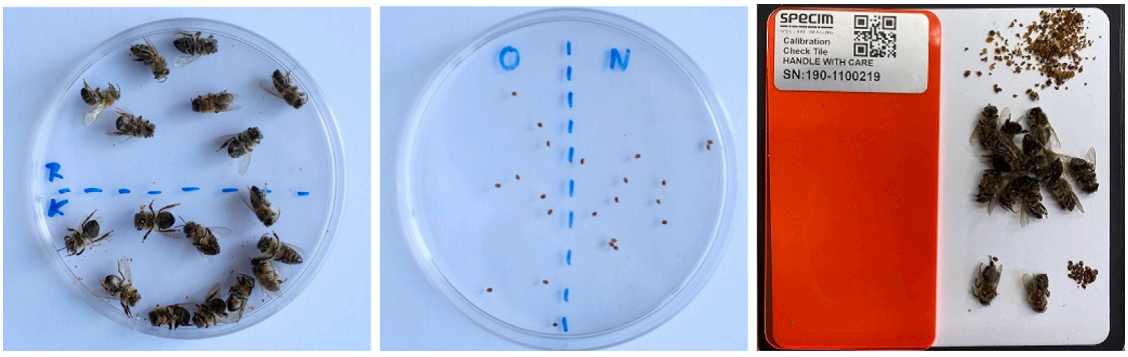
The samples were illuminated with a multispectral unit consisting of 29 LEDs with individually controllable drivers. The LEDs for the unit were selected to cover the spectral range of the experiments. Dedicated software can adjust the illumination to optimize the lighting of the samples. The particular settings of the illumination unit as measured by a fiber optic spectroscope are in Fig. 2(c). Each image in the dataset includes a spectral calibration target with known spectral properties that allow for rectification of the spectra as shown in Fig. 2(b).



(a) Bees mixed-up with detritus.

(b) Clockwise: Bees; detritus; Varroa mites.

(c) Sample with detritus, bees, bees with Varroa mites and Varroa mites on the calibration plate.



(d) Upper part: freshly collected bees; Lower part: 3 months old bee samples.

(e) Left part: 3 month old Varroa mites; Right part: freshly collected Varroa mites.

(f) Sample with clustered detritus, bees, bees with Varroa mites and Varroa mites on the calibration plate.

Fig. 1. Samples from the HS dataset utilized in calibration (a,b, d, e) and testing (c, f). The images are taken with a digital camera and they are not RGB visualizations of the hyperspectral images.

(a) Used Specim IQ camera.

(b) Sample on the calibration plate.

(c) Illumination unit spectrum.

Fig. 2. Utilized camera and measurement setup. The images are taken with a digital camera and they are not RGB visualizations of the hyperspectral images.

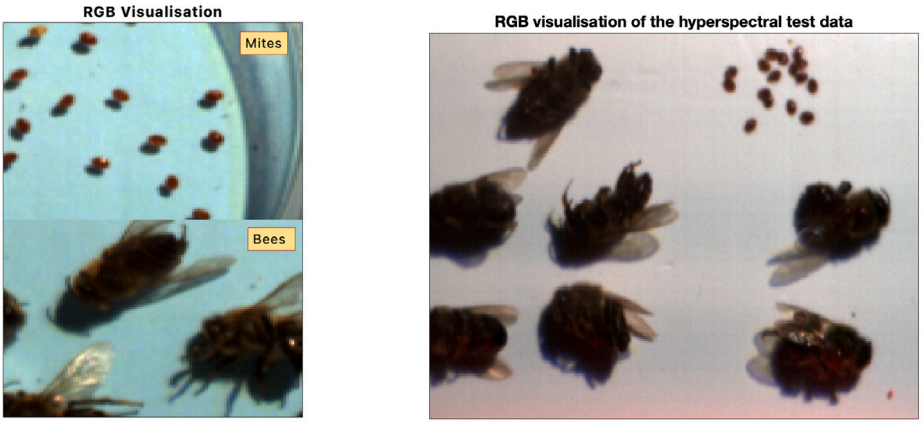


Fig. 3. RGB visualizations of the hyperspectral images with 204 bands for the (a) model calibration data and (b) model testing data. The images are taken with the hyperspectral camera in Fig. 2(a).

For the modeling goal, a group of images from the dataset was selected, containing two sets of images of dead bees and Varroa mites collected from bee detritus, a total of six images. The first image set was used to calibrate the models, whereas the second was held out for testing purposes.

Fig. 3 showcases RGB visualizations of the 204-wavelength images utilized for developing the model. The calibration set in Fig. 3(a) has separate bees and Varroa mites placed on a petri dish on top of a white paper sheet. The testing hyperspectral image is present in Fig. 3(b) and presents bees (left), Varroa mites (up, right), and bees that have Varroa mites on top of them (down, right). The goal of the model is not only to discriminate between the bees and Varroa mites but also to correctly identify the Varroa mites on top of a bee - a more likely scenario in real-case studies. Another notable difference between the calibration and testing sets is the different backgrounds, which can also differ in real-case scenarios.

2.3. Mathematical methods and workflow

This subsection describes the spectral reconstruction and clustering methods, followed by the wavelength selection technique and the algorithms used.

2.3.1. Spectral reconstruction and clustering

The present section presents the mathematical methods building blocks for the workflow in Fig. 4. The first procedure is to center and scale the spectra by subtracting the wavelength mean and dividing by its standard deviation, as seen in Eq (1), where $\mathbf{x}_i$ is an individual wavelength i, $\hat{x}_i$ is the mean of the wavelength and σ_i its standard deviation. Centering ensures that the largest variation profile does not mimic the spectral average and that principal components (PCs) pass through the origin, and scaling ensures that wavelengths are given the same importance in the model.

$$\mathbf{x}_i = \frac{\mathbf{x}_{raw,i} - \hat{x}_i}{\sigma_i} \tag{1}$$

In PCA, the scaled and centered spectra are decomposed into spectral profiles named Principal Components (PCs). The method can be summarized by Eq. (2), where $\mathbf{T}$ is the score matrix, $\mathbf{P}$ is the loadings matrix and $\mathbf{E}$ is the residual matrix.

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \tag{2}$$

The reconstruction of spectra is made only with the PCs whose scores $\mathbf{t}$ have a high absolute correlation with the discriminating variable $\mathbf{y}$. This ensures that variational profiles that discriminate between background and insects or principal components related to noise are not included. The pixels that are Varroa mites or bees are extracted from the calibration data based on the calibration set ground truth. These pixels' scores are then evaluated in relation to a dummy variable with two possible values: '1' in case of membership to the 'bee' class and '0' in case of membership to the 'mite' class. The absolute correlation coefficients are calculated as in Eq. (3), where $\mathbf{t}_i$ is the score vector of selected pixels for the ith PC.

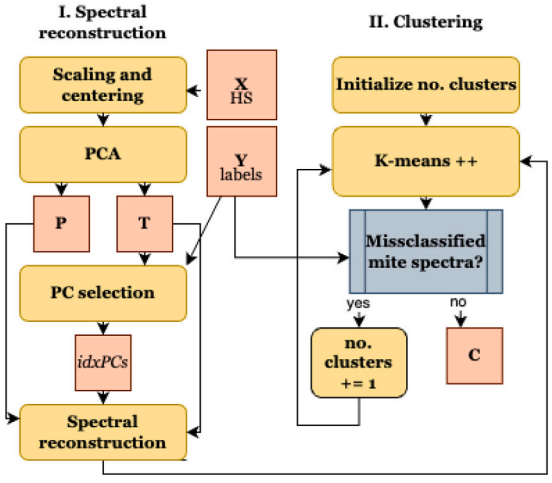


Fig. 4. Cluster formation workflow.

$$\rho(\mathbf{t}_i, \mathbf{y}) = \frac{cov(\mathbf{t}_i, \mathbf{y})}{\sigma_{t_i}\sigma_{\mathbf{y}}} \tag{3}$$

The reconstructed spectra are then subjected to a modified K-means++ algorithm with the methodology presented in Duma et al. (2023a). The initial clustering is considered correct if there are no 'mite' pixels miss-classified, meaning there are no false alarms, pixels inside bees classified as Varroa mites, or un-detected Varroa mites. If one or the previous cases occurs, the number of clusters is increased until the criteria are met.

To utilize the calibrated clusters $\mathbf{C}$ with a new image, the spectra need to be scaled, centered using the calibration mean and center, and then projected into the PCA model, as in the following set of equations:

$$\begin{aligned}\mathbf{X}_{new} &= \frac{\mathbf{X}_{raw,new} - \hat{\mathbf{x}}}{\sigma} \\ \mathbf{T}_{new} &= \mathbf{X}_{new}\mathbf{P} \\ \hat{\mathbf{X}}_{new} &= \mathbf{T}_{new,sel}\mathbf{P}_{sel}^{T}\end{aligned} \tag{4}$$

The Euclidean distance of the newly reconstructed spectra $\hat{\mathbf{X}}_{new}$ is then evaluated to the calibration centroids $\mathbf{C}$, and pixel membership to a cluster is assigned.

In the case of unsupervised clustering, K-means++ was chosen for its ease of use. While K-means is a performant method for unsupervised clustering, it needs to be provided with spectral profiles for efficient cluster identification, as per the methodology presented below; otherwise, it is affected by background variation, shadows, or noise.

However, if one has access to at least a set of labeled data and would like to input the spectra as it is, without spectral profiling, an alternative would be the usage of a multivariate statistical method with discriminant analysis, such as the Kernel Flows-Partial Least-Square (KF-PLS) (Duma et al., 2023b) with Discriminant Analysis. KF-PLS is a variant of optimized Kernel PLS (Rosipal and Trejo, 2001), where the wavelengths are projected via a Kernel function (Gaussian, Laplacian, Matern or Cauchy) into a Reproducing Kernel Hilbert Space via a Kernel Trick (Wu et al., 2005). If the relationship between the spectra and the response variable (classes) is non-linear in the original space, a higher dimension is found where the relationship spectra-classes these becomes linear (Cook and Forzani, 2021). The optimization is based on Kernel Flows (Owhadi and Yoo, 2019) that learns the kernel parameters in a cross-validation manner. The Kernel Flows version of K-PLS is an appropriate choice for the bee-mite discrimination case because (a) it is capable of self-learning the kernel parameters to suit the case, and (b) it can utilize as little information as possible, that includes the necessary information for discrimination, being based in multivariate statistical methods. The workflow of KF-PLS is presented in Fig. 5.

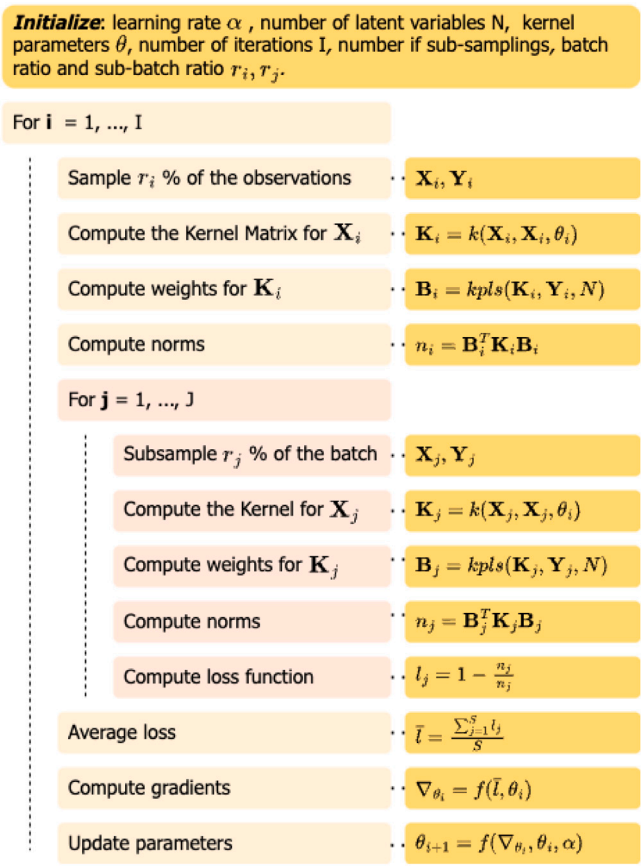


Fig. 5. KF-PLS workflow.

2.3.2. Wavelength selection

Two types of methods based on Partial Least-Squares (PLS) regression have been considered for the wavelength selection. In this study, the SIMPLS version of PLS was utilized (De Jong, 1993). In PLS, the response matrix $\mathbf{Y}$ is also decomposed, as in Eq. (5), where $\mathbf{U}$ is the $\mathbf{Y}$-side score matrix, $\mathbf{Q}$ is the $\mathbf{Y}$-side loadings matrix and $\mathbf{F}$ is the $\mathbf{Y}$-side residual matrix. In this scenario, the PLS version utilized is PLS with Discriminant Analysis (PLS-DA), as the $\mathbf{y}$ variables give information on the membership or non-membership of a pixel to a cluster.

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^{T} + \mathbf{F} \tag{5}$$

The first wavelength selection method is the iterative forward selection based on the PLS model between the wavelengths and the group membership variable $\mathbf{y}$. The variable selection method is a modified version of the Forward Interval PLS (FiPLS) (Balabin and Smirnov, 2011), where instead of utilizing intervals and cross-validation, the search was exhaustive throughout all variables. To initialize the selection of variables, the first three correlated variables can be calculated with the correlation coefficient with respect to the response variable, as in Eq. (3).

The regression coefficients can be obtained through Eq. (6), where $\mathbf{W}$ is the matrix of rotated loadings in the direction of maximum $\mathbf{X}$ and $\mathbf{y}$ covariance.

$$\mathbf{b} = \mathbf{W}(\mathbf{P}^{T}\mathbf{W})^{-1}\mathbf{Q}^{T} \tag{6}$$

After the initial selection, the following steps can be followed, as seen in **Algorithm** 1: (*i*) add the variables one at a time to the selected list of variables, (*ii*) perform PLS between the selected variables and the response, (*iii*) evaluate which iteration had the highest explained variance of the $\mathbf{y}$-variable R^2, (*iv*) permanently add to the selected list of variables the wavelength whose addition yielded the highest R^2, (*v*) repeat Step *i* until the desired number of variables has been collected. In the present case, adding variables stops when the success conditions of the clustering algorithm are met: no 'mite' pixels are misclassified.

Algorithm 1 R^2-based Variable Selection

Input: spectral matrix, with wavelengths as columns and bee-mite pixels as rows ($\mathbf{X}$), bee-mite discriminating vector ($\mathbf{y}$), the desired number of variables to be selected (V).
Output: vector with selected wavelength indices of ($\mathbf{s}$).
Initialize: initial vector of selected variables ($\mathbf{s}_0$), initial vector of unselected variables ($\mathbf{sn}_0$), initial number of selected (N) and unselected (M) wavelengths.

for $v \leftarrow 1$ to V **do** ▷ Loop until desired number of variables selected.
  $TSS \leftarrow \sum(\mathbf{y} - \bar{\mathbf{y}})^2$ ▷ Calculate the total sum of squares.
  for $m \leftarrow 1$ to M **do** ▷ Loop through unselected variables.
    $\mathbf{s}_{temp} \leftarrow [\mathbf{s}, \mathbf{sn}_m]$ ▷ Add the current unselected variable to the temporary variable list.
    $\mathbf{b} \leftarrow pls(\mathbf{X}_{s_{temp}}, \mathbf{y})$ ▷ Calibrate PLS model.
    $\hat{\mathbf{y}} \leftarrow \mathbf{X}_{s_{temp}}\mathbf{b}$ ▷ Estimate $\mathbf{y}$ from the model.
    $RSS_m \leftarrow \sum(\mathbf{y} - \hat{\mathbf{y}})^2$ ▷ Calculate the prediction sum of squares.
    $R^2_m \leftarrow 1 - \frac{RSS_m}{TSS}$ ▷ Calculate the prediction R^2 for the iteration.
  end for
  $idx \leftarrow max(\mathbf{R}^2)$ ▷ Select the iteration with the highest increase.
  $\mathbf{s}_{N+1} \leftarrow \mathbf{sn}_{idx}$ ▷ Add variable to selected variables.
  $N \leftarrow N + 1$ ▷ Increase the counter of selected variables.
end for

The second version of wavelength variable selection proposed is a modified version of the Covariance Procedures (COVPROC) (Reinikainen and Höskuldsson, 2007), that is presented in **Algorithm 2**. An optional step of relevance to the present paper is sorting the rounds based on their included number of variables.

Algorithm 2 COVPROC-based Variable Selection

Input: spectral matrix, with wavelengths as columns and bee-mite pixels as rows (**X**), bee-mite discriminating vector (**y**), the desired number of COVPROC rounds (R), the total number of wavelengths (I).
Output: selected list of variables (**s**)
Initialize: a vector with '0' values of length I (**a**)

for $r \leftarrow 1$ to R **do**
  $\mathbf{w} \leftarrow pls(\mathbf{X}, \mathbf{y})$ ▷ Extract PLS weights for one latent variable.
  $\mathbf{idxw} \leftarrow sort(|\mathbf{w}|)$ ▷ Sort the weights in descending order and save indexes.
  $\mathbf{w}_r \leftarrow \mathbf{a}$ ▷ Reset $\mathbf{w}_b$
  for $i \leftarrow 1$ to I **do**
    $ii \leftarrow idxw_i$ ▷ Select variable.
    $\mathbf{w}_{r,ii} \leftarrow \mathbf{y}^T\mathbf{x}_{ii}$ ▷ Populate $\mathbf{w}_r$ with selected variable/response covariance.
    $\mathbf{t}_{r,i} \leftarrow \mathbf{X}\mathbf{w}_r$ ▷ Compute iteration's scores.
    $\alpha_i \leftarrow \frac{|\mathbf{y}^T\mathbf{t}_{r,i}|}{\mathbf{t}_{r,i}^T\mathbf{t}_{r,i}}$ ▷ Calculate evaluation metric.
  end for
  $n \leftarrow max(\boldsymbol{\alpha})$ ▷ Extract the index of the maximum $\boldsymbol{\alpha}$.
  $\mathbf{s}_r \leftarrow \mathbf{idxw}_{1,\ldots,n}$ ▷ Extract the variables of the round.
  $\mathbf{s} \leftarrow [\mathbf{s}, \mathbf{s}_r]$ ▷ Append the selected variable list of round r.
  $\mathbf{p}_r \leftarrow \frac{\mathbf{X}^T\mathbf{t}_{r,n}}{\mathbf{t}_{r,n}^T\mathbf{t}_{r,n}}$ ▷ Compute the loadings of the round r.
  $\mathbf{X} \leftarrow \mathbf{X} - \mathbf{t}_{r,n}\mathbf{p}_r^T$ ▷ Deflate **X**.
end for

3. Results and discussion

3.1. Variation profiles

After centering and scaling the image, the principal component analysis revealed that the background of the hyperspectral images represents 97% of the systematic variation, as seen in Fig. 6(a) and Fig. 6(c). One can observe that the 2nd and 3rd principal components, responsible for capturing the systematic variation between bees and Varroa mites, can be utilized for spectral reconstruction. The two PCs sum up to approximately 2% of the total variation. These were also the principal components whose scores had the highest absolute correlation coefficient with the bee-mite discriminating variable and thus were selected for reconstruction.

The loadings of the discriminating PCs are displayed in Fig. 6(b). It is observable that the 2nd principal component has high loadings at the beginning and end of the spectrum, whereas the second one describes the middle part of the spectra. The higher the absolute value of a wavelength loading to a principal component, the more important it is for the variation profile. The rest of the principal components seem to either explain variation related to wings (PC4), shadows (PC4 to PC8), or noise (over PC9).

3.2. Full-spectral clustering

A minimum of four clusters are necessary for the k-means algorithm to separate the Varroa mites from the bees, as seen in Fig. 7(a). If the clustering is ran with a lower number of clusters, the Varroa mite and some parts of the bee are being clustered together.

The full-wavelength model utilizes the spectral reconstruction of the second and third principal components as an input to the k-means algorithm. The testing image (Fig. 7(b)) has been centered and scaled with the calibration image mean and standard deviation prior to the PCA projection and spectral reconstruction.

Furthermore, Fig. 7 shows that with four clusters, the bees are not fully separated from the background, yet the Varroa mites are already separated into their cluster. The mites can even be detected on top of the bees Fig. 7(b).

The KF-PLS algorithm needed 150 iterations to converge. It was trained on four classes containing 300 pixels from each category: Varroa mite, bee wing, bee body and background. The results of the KF-PLS on the test image are seen in Fig. 8. The hyperparameters utilized to obtain the results were a learning rate of 0.1, Polyak's momentum for parameter update, 20 sub-samplings per iteration, and a batch sampling ratio of 50% of the observations for each iteration. The Kernel function utilized was Matern5/2. As shown in Fig. 8(b) the KF-PLS needed 6 latent variables to converge, which was confirmed in the external loop evaluated with the optimized learned kernel parameters.

3.3. Wavelength selection

Due to the inconsistent variation of the last 10 wavelengths, they have not been included in the list of selected variables, regardless of the wavelength selection method. The variable's value varied more with the spatial location of the pixel and not the type of material measured.

In the R^2 variable selection method, the initializing step requires analyzing the correlation coefficients of individual wavelengths with the response variable. The absolute value of the correlation coefficients is showcased in Fig. 9(a). The higher the wavelength, the higher the absolute correlation with the discriminating variables. The three more-correlated variables were included in the initialization vector for the R^2-based selection.

Even though the maximum R^2 value of 86% was obtained in a PLS model with approx. 40 variables (Fig. 9(b), only the first 12 selected variables were necessary to obtain the test partition discrimination correctly. The 12 essential wavelengths can be observed in Fig. 9(c). A notable selection is the one at 800 nm, which will be important in the further discussed results.

As observed in Fig. 9(d), Varroa mites could be identified, even if they were placed on top of the bees, but two false alarms were observed as well: one in the wing of the upmost bee and another one in the leg of another bee. These errors are presented as singular pixels (Fig. 9(e)) and can be processed, if needed, with image processing techniques such as morphological opening (Fig. 9(f)). The size of the structuring element in this case was two pixels, both for the erosion and dilation in the opening process. The opening operation is successful in eliminating false alarms if the false positives are smaller in size than the true positives.

The second variable selection method compared, COVPROC, has the variable selection done in rounds. Fig. 10(b) shows the evaluation metric α for each of the rounds and the number of variables necessary for the peak of the evaluation metric.

As seen in Fig. 10(a), two variables that were in the exclusion area were selected in the first round. In the second round, a non-intuitive list of 28 variables from the beginning of the spectrum was selected. The spectral region selected presents the highest overlaps in the distributions of bees and Varroa mites. In round three, only one variable, around 800 nm, was responsible for reaching the maximal evaluation metric. In contrast, round 4 presents an exhaustive list of 157 variables, indicating that further evaluations should not be considered.

With the 1st round excluded, the variables selected in the 2nd and the 3rd round were enough to yield a good classification. It was observed practically that adding the 3rd round before the 2nd gave even better results. Instead of the 29 necessary variables in the round 2+3 succession, only four variables were needed if the selection order was round 3+2. This resulted in the image Fig. 10(c) being obtained with the spectral reconstruction of only four wavelengths, one around 800 nm and three consecutive wavelengths around 500 nm.

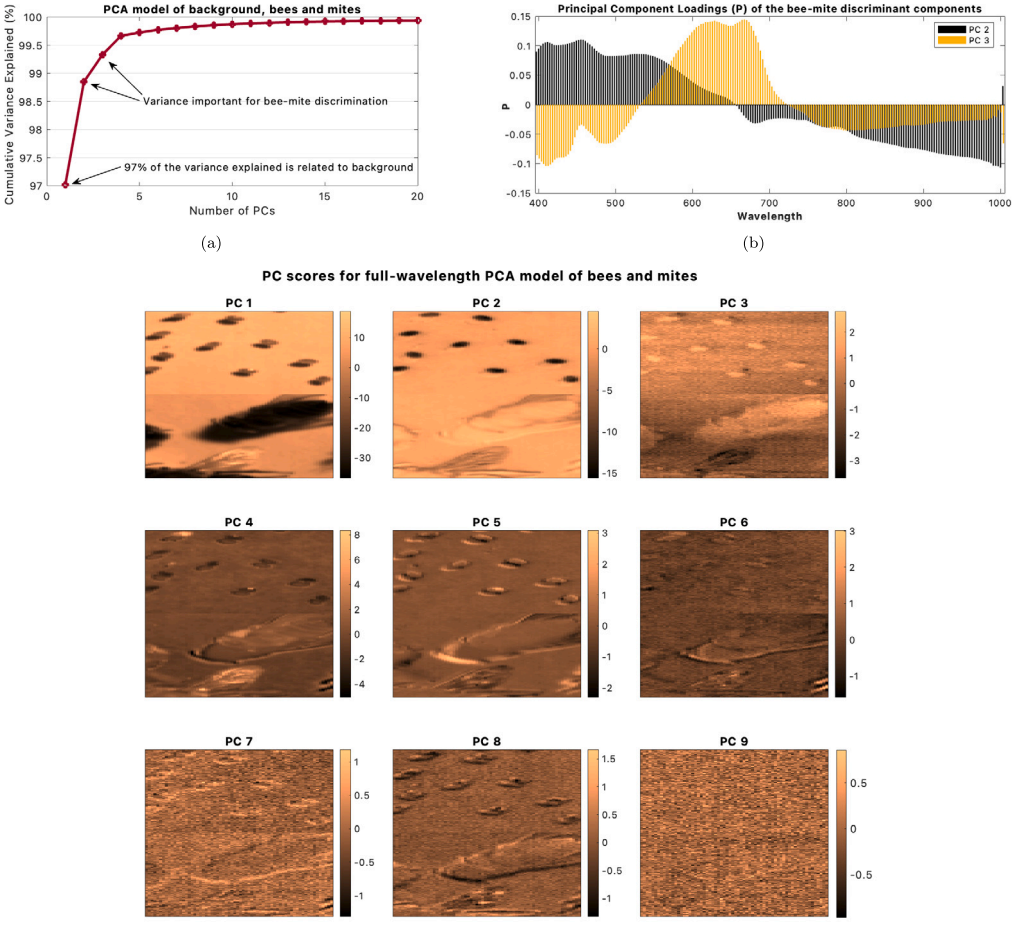


Fig. 6. (a) Explained variance of calibration dataset for the principal components model. (b) Score values for the first 10 PCs. The PCs responsible for distinguishing between the bees and Varroa mites are PC1 and PC2, and the loadings of the discriminant PCs (c).

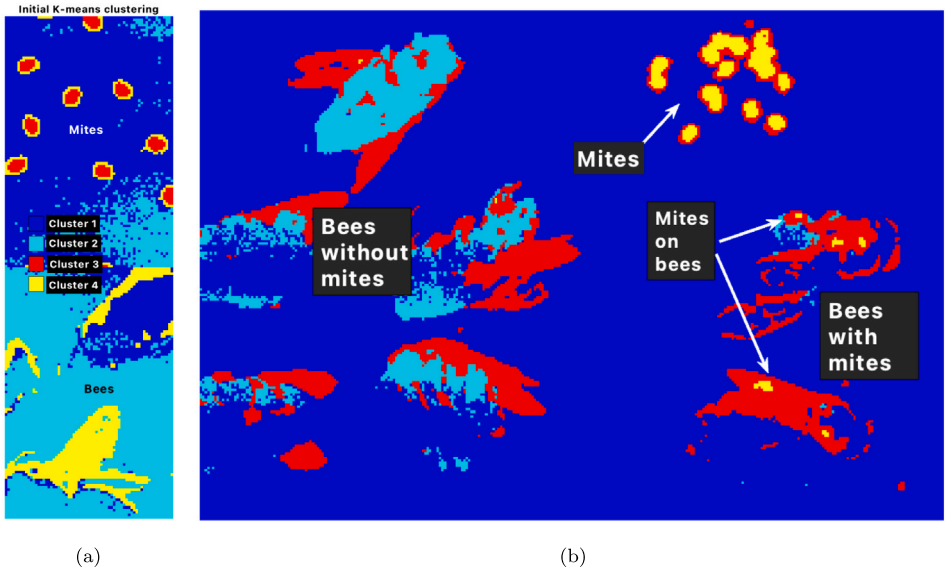


Fig. 7. (a) Calibration and (b) testing image k-means clustering results in four clusters on full reconstructed spectra (Note: Cluster colors are not consistent between plots, red and yellow have been switched for visibility in the right-hand plot).

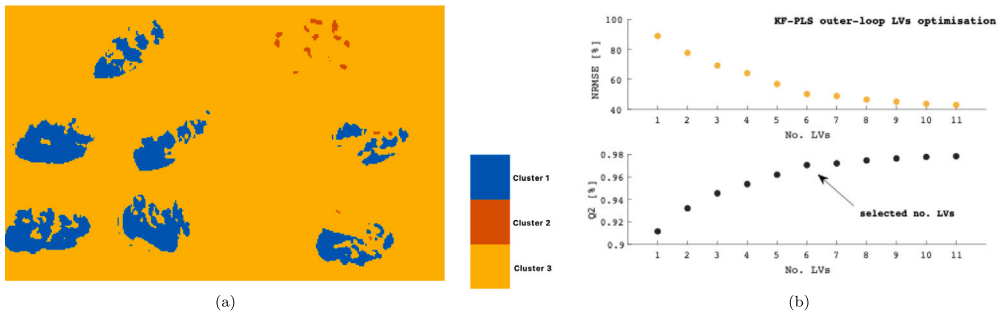


Fig. 8. The KF-PLS results for a Matern5/2 Kernel function on the test image (a). The number of optimal LVs in the KF-PLS was 6. After the optimal number of LVs, the identification of Varroa mites was still successful. (Note: Cluster colors are chosen to enhance visibility and are not consistent between methods.)

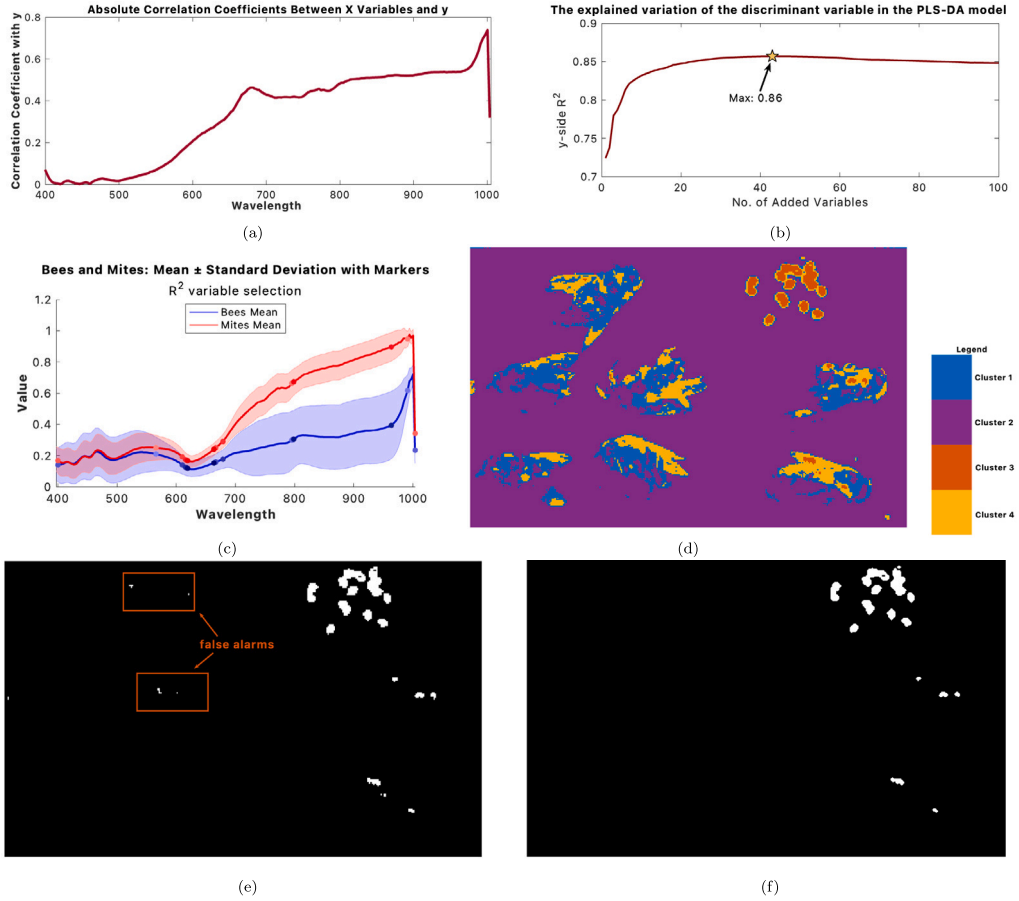


Fig. 9. (a) The absolute correlation coefficients between wavelengths and y-variable — the initializing step of the PLS-DA model, whose explained variance increases with the number of added variables is present in figure (b). (c) showcases the minimum number of variables to correctly identify the independent data set (d). The mask of the mite cluster is presented in (e) and the result after the opening morphological operation on the mask is presented in (f). (Note: Cluster colors are chosen to enhance visibility and are not consistent between methods.)

3.4. Discussion on future research

The present study presents for the first time the utilizing of hyperspectral images to detect Varroa mites on bees' bodies. Studies with multispectral cameras have been done in literature, such as Bjerge et al. (2019), where the camera had 19 spectral bands. In the current study, 204 bands were utilized, a tenfold increase. This provides more robustness to the analysis and provides more information on what wavelengths would best discriminate between bees and mites. These wavelengths were identified to be 492.97 nm, 498.8 nm, 507.56 nm and 796.74 nm. These are similar to bands 470 nm, 630 nm and 780 nm identified in Bjerge et al. (2019). However, based on the HS data, the 630 nm band gives inconsistent results, which can lead to false alarms, confusing pollinated bee's legs as mites.

Precise identification of discriminating wavelengths is crucial for developing a real-time monitoring system in which bees enter the hive through a limited passway, as suggested in Bjerge et al. (2019), which is continuously captured by a camera. With suitable narrow-spectrum illumination and a common camera with a custom filter, Varroa mite-infested bees could be easily distinguished without the need for an expensive multispectral camera. Two possible setups with one and two cameras placed above a tunnel in front of the beehive entrance are shown in Fig. 11 The single camera approach is inspired by our bee counting setup presented in Bilik et al. (2024a) and the two camera

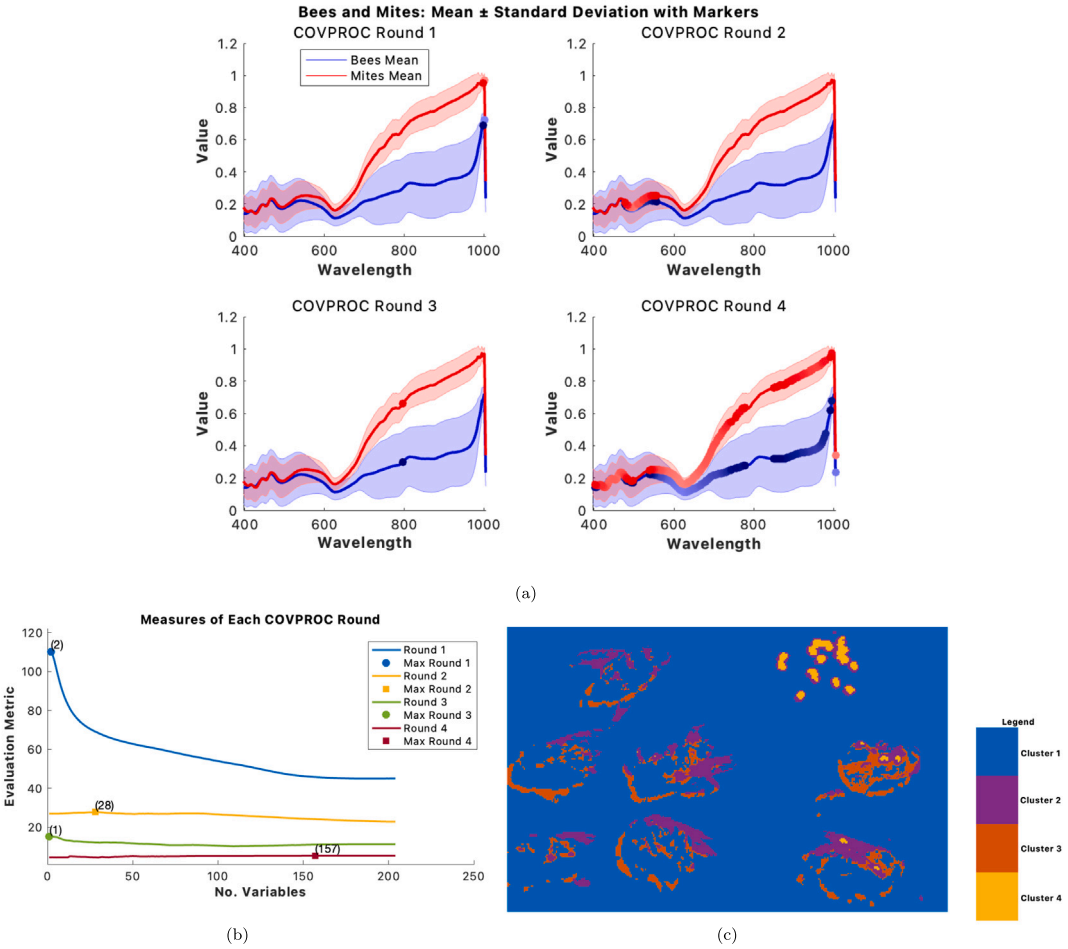


Fig. 10. (a) The variables selected in each COVPROC round. A light color in the marker indicates that the variable was selected early in the selection process. A darker marker indicates that the variable was selected later in the round. The evaluation metric results for each round are found in Fig. (b). (Note: Cluster colors are chosen to enhance visibility and are not consistent between methods.)

approach was inspired by Bjerge et al. (2019). The main advantages of using one camera are smaller dimensions and higher simplicity due to fewer optical and electronic elements. On the other hand, the two-camera approach brings the benefit of imaging the bees from both sides and despite the higher requirements for synchronization and performance, it could bring more accurate results in the Varroa mite detection. The two-camera approach also complicates the separation of the background — the cameras effectively point at each other. Although the restricted hive access might be considered as controversial, on experience from our previous measurements presented in Nevlacil et al. (2023) and Bilik et al. (2024a), bees are getting used to the restricted hive access in several days with the benefit of much more stable imaging with constant illumination and background.

Such a device could be used for long-term and regular beehive monitoring, which could be more reliable in Varroa mite detection than broad-spectrum white illumination. Sensing in such illumination could be challenging due to high color similarity between the mite and bee or due to partial occlusion of the mite between the bee's abdomen segments. In future work, we plan to improve and test our device under development described in Bilik et al. (2021b) based on the findings of this study.

The current study considered only discrimination between bees and mites, as that would be the main application. Thus, the rest of the hive detritus has been left unstudied. The HS images could reveal more information from this fraction as well. Our proposed approach could be used for the detritus analysis, including Varroa mites detection as suggested in Picek et al. (2022). This would require further development of the method to distinguish between other objects in detritus, such as wax, not just the bees and Varroa mites. However, in the proposed use case, the bees entering the hive should be groomed and thus not carrying any wax. Nevertheless, further methodology development is planned to analyze the next in-field measurements.

The usage of the method is not limited to bee-parasite identification. The application can be easily transferred to hyperspectral discrimination cases for example in waste management (Tao et al., 2023), or food control (Liu et al., 2024a,b), where existing methodologies relying on deep learning require an extensive number of samples. The deep learning approach can efficiently solve more complex problems with more classes to discriminate or with a higher similarity between the target classes. The present study gives an alternative for when (a) the available number of samples is low (b) the discrimination problem is not so complex.

4. Conclusion

The present article demonstrated that hyperspectral imagery can be utilized in beehive health monitoring by detecting Varroa mites on bees. The research revealed that unsupervised (K-means) and supervised (KF-PLS) clustering methods are efficient in identifying the parasites from the hyperspectral images. All parasites have been correctly identified. However, the unsupervised methods required more preprocessing of the data than the supervised methods. This is due to the background of the image dominating the scene, so the methods have a tendency to discriminate between the background and

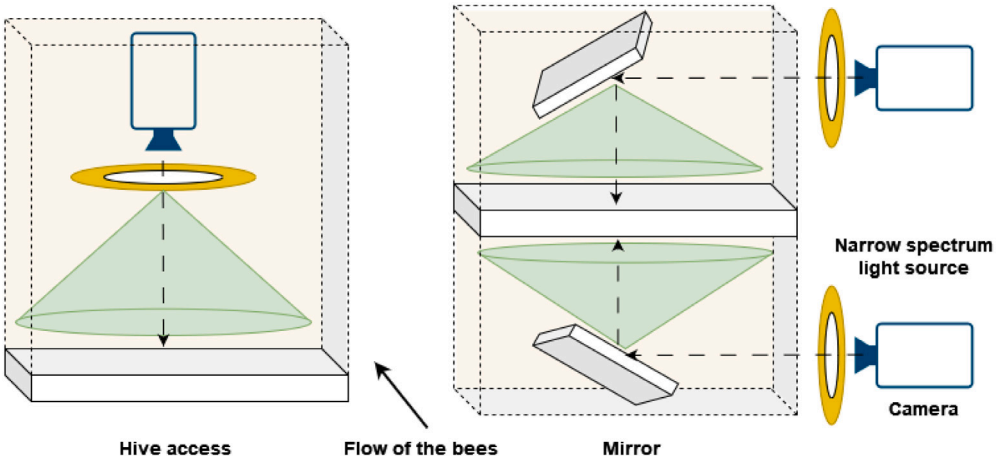


Fig. 11. Proposed detection setup based on the restricted hive access with one (left) and two (right) cameras.

anything else, thus clustering the bees and mites together. This can be alleviated by reassembling the spectra from its most discriminant components. The unsupervised method (KF-PLS) proved efficient for bee-mite discrimination without significant preprocessing.

For potential online monitoring of beehives, it is more realistic to utilize normal cameras with filters and custom light sources, utilizing only wavelengths that inform about the difference between bees and mites. Two methods for finding the wavelengths that best discriminate between bees and mites are the coefficient of determination (R^2)-based method and the COVPROC method. Both proved high wavelength reduction from the initial 204-wavelength set: the COVPROC method reduced the wavelengths to 4 essential ones, whereas the R^2 method resulted in 12 wavelengths selected that were able to give a good discrimination in the unsupervised methodology.

The current HS dataset has been made publicly available to support further research in this field. In future research, the dataset will be expanded with more in-field measurements. The methodology will also be expanded to distinguish between more classes of bee detritus (not only bees and Varroa mites but also wax, pollen, sugar, and other debris). Furthermore, a deeper analysis of the methods' discrimination ability on a more complex scene is needed. The results of discriminating wavelengths can already be applied to the real-time monitoring setup.

CRediT authorship contribution statement

Zina-Sabrina Duma: Writing – original draft, Visualization, Validation, Software, Methodology, Data curation, Conceptualization. **Tomas Zemcik:** Writing – original draft, Visualization, Resources, Investigation, Data curation. **Simon Bilik:** Writing – original draft, Visualization, Resources, Project administration, Investigation, Data curation, Conceptualization. **Tuomas Sihvonen:** Writing – original draft, Validation, Software, Methodology, Formal analysis, Conceptualization. **Peter Honec:** Supervision, Resources, Project administration. **Satu-Pia Reinikainen:** Supervision, Resources, Project administration, Formal analysis, Conceptualization. **Karel Horak:** Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

We have shared the DOI to the dataset, which was uploaded on Kaggle" to "A hyperspectral frame from each setup is available in Kaggle at http://dx.doi.org/10.34740/KAGGLE/DSV/7845514, URL https://www.kaggle.com/dsv/7845514.

Acknowledgments

Funding from Research Council of Finland for Centre of Excellence of Inverse Modeling and Imaging, Finland, project number 353095, is acknowledged, and the Research Council of Finland through the Flagship of Advanced Mathematics for Sensing, Imaging and Modeling, Finland (decision number 359183) is also acknowledged. The presented project was further supported by the grant no. FEKT-S-23-8451 "Research on advanced methods and technologies in cybernetics, robotics, artificial intelligence, automation and measurement" from the Internal science fund of Brno University of Technology, Czech Republic.

References

Balabin, R.M., Smirnov, S.V., 2011. Variable selection in near-infrared spectroscopy: benchmarking of feature selection methods on biodiesel data. Anal. Chim. Acta 692 (1–2), 63–72.

Behmann, J., Acebron, K., Emin, D., Bennertz, S., Matsubara, S., Thomas, S., Bohnenkamp, D., Kuska, M.T., Jussila, J., Salo, H., Mahlein, A.-K., Rascher, U., 2018. Specim IQ: Evaluation of a new, miniaturized handheld hyperspectral camera and its application for plant phenotyping and disease detection. Sensors 18 (2), http://dx.doi.org/10.3390/s18020441, URL https://www.mdpi.com/1424-8220/18/2/441.

Bilik, S., Janakova, I., Ligocki, A., Ficek, D., Horak, K., 2024a. Computer vision approaches for automated bee counting application. arXiv:2406.08898.

Bilik, S., Kratochvila, L., Ligocki, A., Bostik, O., Zemcik, T., Hybl, M., Horak, K., Zalud, L., 2021a. Visual diagnosis of the varroa destructor parasitic mite in honeybees using object detector techniques. Sensors 21 (8), 2764.

Bilik, S., Ligocki, A., Nevlacil, J., 2021b. Bee health monitor. URL https://github.com/boortel/Bee-Health-Monitor, Open source software available from https://github.com/boortel/Bee-Health-Monitor.

Bilik, S., Zemcik, T., Duma, Z.-S., Sihvonen, T., Honec, P., Reinikainen, S.-P., Horak, K., 2024b. Bee Dataset BUT-HS. Kaggle, http://dx.doi.org/10.34740/KAGGLE/DSV/7845514, URL https://www.kaggle.com/dsv/7845514.

Bilik, S., Zemcik, T., Kratochvila, L., Ricanek, D., Richter, M., Zambanini, S., Horak, K., 2024c. Machine learning and computer vision techniques in continuous beehive monitoring applications: A survey. Comput. Electron. Agric. 217, 108560. http://dx.doi.org/10.1016/j.compag.2023.108560, URL https://www.sciencedirect.com/science/article/pii/S0168169923009481.

Bjerge, K., Frigaard, C.E., Mikkelsen, P.H., Nielsen, T.H., Misbih, M., Kryger, P., 2019. A computer vision system to monitor the infestation level of varroa destructor in a honeybee colony. Comput. Electron. Agric. 164, 104898. http://dx.doi.org/10.1016/j.compag.2019.104898, URL https://www.sciencedirect.com/science/article/pii/S0168169918310329.

Bjorgan, A., Randeberg, L.L., 2015. Real-time noise removal for line-scanning hyperspectral devices using a minimum noise fraction-based approach. Sensors 15 (2), 3362–3378.

Cook, R.D., Forzani, L., 2021. PLS regression algorithms in the presence of nonlinearity. Chemometr. Intell. Lab. Syst. 213, 104307.

De Jong, S., 1993. SIMPLS: An alternative approach to partial least squares regression. Chemometr. Intell. Lab. Syst. 18 (3), 251–263.

Duma, Z.-S., Sihvonen, T., Härmä, P., Reinikainen, S.-P., 2023a. Colorimetric similarity evaluation methodology for heterogeneous rock surfaces using digital imaging. J. Cult. Herit. 64, 244–254.

Duma, Z.-S., Susiluoto, J., Lamminpää, O., Sihvonen, T., Reinikainen, S.-P., Haario, H., 2023b. KF-PLS: Optimizing kernel partial least-squares (K-PLS) with kernel flows. arXiv preprint arXiv:2312.06547.

Eliash, N., Mikheyev, A., 2020. Varroa mite evolution: A neglected aspect of worldwide bee collapses? Curr. Opin. Insect Sci. 39, 21–26.

Flores, J.M., Gámiz, V., Jiménez-Marín, Á., Flores-Cortés, A., Gil-Lebrero, S., Garrido, J.J., Hernando, M.D., 2021. Impact of varroa destructor and associated pathologies on the colony collapse disorder affecting honey bees. Res. Vet. Sci. 135, 85–95.

Hall, H., Bencsik, M., Newton, M., 2023. Automated, non-invasive varroa mite detection by vibrational measurements of gait combined with machine learning. Sci. Rep. 13 (1), 10202.

Hämäläinen, J., Kärkkäinen, T., Rossi, T., 2020. Improving scalable K-means++. Algorithms 14 (1), 6.

Ikäheimo, E., Jussila, J., 2018. Introducing Specim IQ. Specim, Spectral Imaging Oy, Finland, URL https://www.specim.com/downloads/iq/Introducing-Specim-IQ.pdf.

Jack, C.J., Ellis, J.D., 2021. Integrated pest management control of Varroa destructor (Acari: Varroidae), the most damaging pest of (Apis mellifera L.(Hymenoptera: Apidae)) colonies. J. Insect Sci. 21 (5), 6.

Khan, A., Vibhute, A.D., Mali, S., Patil, C., 2022. A systematic review on hyperspectral imaging technology with a machine and deep learning methodology for agricultural applications. Ecol. Inform. 69, 101678.

König, A., 2020. VarroaCounter–towards automating the varroa screening for alleviated bee hive treatment. In: SEIA'2019 Conference Proceedings. pp. 244–247.

Liu, M., Cui, M., Xu, B., Liu, Z., Li, Z., Chu, Z., Zhang, X., Liu, G., Xu, X., Yan, Y., 2023. Detection of varroa destructor infestation of honeybees based on segmentation and object detection convolutional neural networks. AgriEngineering 5 (4), 1644–1662.

Liu, Y., Feng, H., Fan, Y., Yue, J., Chen, R., Ma, Y., Bian, M., Yang, G., 2024a. Improving potato above ground biomass estimation combining hyperspectral data and harmonic decomposition techniques. Comput. Electron. Agric. 218, 108699.

Liu, Y., Feng, H., Yue, J., Jin, X., Fan, Y., Chen, R., Bian, M., Ma, Y., Li, J., Xu, B., et al., 2024b. Improving potato AGB estimation to mitigate phenological stage impacts through depth features from hyperspectral data. Comput. Electron. Agric. 219, 108808.

Lu, B., Dao, P.D., Liu, J., He, Y., Shang, J., 2020. Recent advances of hyperspectral imaging technology and applications in agriculture. Remote Sens. 12 (16), 2659.

Månefjord, H., Müller, L., Li, M., Salvador, J., Blomqvist, S., Runemark, A., Kirkeby, C., Ignell, R., Bood, J., Brydegaard, M., 2022. 3D-printed fluorescence hyperspectral lidar for monitoring tagged insects. IEEE J. Sel. Top. Quantum Electron. 28 (5: Lidars and Photonic Radars), 1–9.

Mekha, P., Teeyasuksaet, N., Sompowloy, T., Osathanunkul, K., 2022. Honey bee sound classification using spectrogram image features. In: 2022 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering. (ECTI DAMT & NCON), IEEE, pp. 205–209.

Nevlacil, J., Bilik, S., Horak, K., 2023. Raspberry pi bee health monitoring device. In: Proceedings I of the 29th Student EEICT 2023. Brno University of Technology, Faculty of Electrical Engineering and Communication, pp. 226–230.

Owhadi, H., Yoo, G.R., 2019. Kernel flows: From learning kernels from data into the abyss. J. Comput. Phys. 389, 22–47.

Picek, L., Novozamsky, A., Frydrychova, R.C., Zitova, B., Mach, P., 2022. Monitoring of varroa infestation rate in beehives: A simple ai approach. In: 2022 IEEE International Conference on Image Processing. ICIP, pp. 3341–3345. http://dx.doi.org/10.1109/ICIP46576.2022.9897809.

Rajula, H.S.R., Verlato, G., Manchia, M., Antonucci, N., Fanos, V., 2020. Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment. Medicina 56 (9), 455.

Reinikainen, S.-P., Höskuldsson, A., 2003. COVPROC method: strategy in modeling dynamic systems. J. Chemometr.: J. Chemometr. Soc. 17 (2), 130–139.

Reinikainen, S.-P., Höskuldsson, A., 2007. Multivariate statistical analysis of a multi-step industrial processes. Anal. Chim. Acta 595 (1–2), 248–256.

Rodarmel, C., Shan, J., 2002. Principal component analysis for hyperspectral image classification. Surv. Land Inf. Sci. 62 (2), 115–122.

Rosipal, R., Trejo, L.J., 2001. Kernel partial least squares regression in reproducing kernel hilbert space. J. Mach. Learn. Res. 2 (Dec), 97–123.

Roth, M.A., Wilson, J.M., Tignor, K.R., Gross, A.D., 2020. Biology and management of varroa destructor (Mesostigmata: Varroidae) in Apis mellifera (Hymenoptera: Apidae) colonies. J. Integr. Pest Manag. 11 (1), 1.

Tao, J., Gu, Y., Hao, X., Liang, R., Wang, B., Cheng, Z., Yan, B., Chen, G., 2023. Combination of hyperspectral imaging and machine learning models for fast characterization and classification of municipal solid waste. Resour., Conserv. Recy. 188, 106731.

Wu, G., Chang, E.Y., Panda, N., 2005. Formulating distance functions via the kernel trick. In: Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining. pp. 703–709.

Yun, Y.-H., Li, H.-D., Deng, B.-C., Cao, D.-S., 2019. An overview of variable selection methods in multivariate analysis of near-infrared spectra. TRAC Trends Anal. Chem. 113, 102–115.

Zemčík, T., Horak, K., 2023. On hyperspectral analysis of water soluble writing inks. In: Proceedings II of the 28th Student EEICT 2023 Selected Papers. pp. 237–242. http://dx.doi.org/10.13164/eeict.2023.237.