Research Square

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

HyperCookies A hyperspectral dataset of cookies for spectral-spatial machine learning

Gerardo Mora Jimena
KU Leuven https://orcid.org/0009-0009-9460-5791
Astrid Tempelaere
KU Leuven https://orcid.org/0000-0003-3807-6695
Junyan Li
KU Leuven https://orcid.org/0009-0000-3266-4619
Mariette Schönfeld
KU Leuven https://orcid.org/0009-0000-3335-6538
Noah Steidle
KU Leuven https://orcid.org/0009-0002-2907-3791
Hendrik Blockeel
KU Leuven https://orcid.org/0000-0003-0378-3699
Bart De Ketelaere
KU Leuven https://orcid.org/0000-0002-5140-1643
Mariya Ishteva
KU Leuven https://orcid.org/0000-0002-7951-536X
Wannes Meert
KU Leuven https://orcid.org/0000-0001-9560-3872
Ward Melis
KU Leuven https://orcid.org/0000-0002-5066-5209
Pieter Verboven
KU Leuven https://orcid.org/0000-0001-9542-8285
Wouter Saeys
wouter.saeys@kuleuven.be

KU Leuven https://orcid.org/0000-0002-5849-4301

Data Note

Keywords:

Posted Date: November 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11633892/v1

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

Additional Declarations: The authors declare no competing interests.

HyperCookies A hyperspectral dataset of cookies for spectral-spatial machine learning

Gerardo Mora Jimena[1#], Astrid Tempelaere[1#], Junyan Li[1], Mariette Schönfeld[2,3], Noah Steidle[4], Hendrik Blockeel[2,3], Bart De Ketelaere[1], Mariya Ishteva[4], Wannes Meert[2,3], Ward Melis[4], Pieter Verboven[1], Wouter Saeys[1*]

[1] KU Leuven, Belgium, Department of Biosystems, MeBioS

[2] KU Leuven, Belgium, Department of Computer Sciences, Declarative Languages and Artificial Intelligence (DTAI)

[3] KU Leuven, Belgium, KU Leuven Institute for AI, Leuven.AI

[4] KU Leuven, Belgium, Department of Computer Sciences, Numerical Analysis and Applied Mathematics (NUMA)

* Corresponding author: wouter.saeys@kuleuven.be

# These authors contributed equally

Abstract

Hyperspectral imaging (HSI) combines spatial and spectral information, producing high-dimensional datasets for which machine learning (ML) methods have become a key tool for classification. However, publicly available HSI datasets that explicitly require the joint exploitation of both spatial and spectral dimensions for classification remain limited. Here, we present HyperCookies, a carefully annotated HSI dataset comprising images of 160 cookies belonging to four classes defined by both spectral and spatial characteristics. Spectral variation is introduced through chocolate droplets with different cocoa contents, and spatial variation is represented by two topping arrangements. HSI data were acquired using two complementary line-scan cameras covering the visible and near-infrared (VNIR, 400–1000 nm) and short-wave infrared (SWIR, 900–1700 nm) ranges. The cookies were imaged both before and after baking, enabling the study of processing-induced changes. Baseline experiments using both spectral ML classifiers and image-based convolutional neural networks demonstrate that neither spectral nor spatial information alone is sufficient to reliably discriminate all four classes. The HyperCookies dataset is therefore made publicly available as a challenging and realistic benchmark for

the development and evaluation of methods that jointly exploit spectral and spatial information in hyperspectral images.

Background & Summary

In computer vision, information is commonly acquired using digital cameras operating in the visible region of the electromagnetic spectrum. Standard color cameras record reflected light in three broad, overlapping wavelength bands, centered in the red (~600-630 nm), green (~530-550 nm), and blue (~430-470 nm) regions (RGB), producing images that approximate human visual perception [1]. The resulting RGB images have enabled the rapid development of computer vision methods for automated and objective image analysis across a wide range of applications [2–4].

However, many materials exhibit spectral characteristics that cannot be fully captured using only three visible wavelength bands. Recent examples in the agrifood domain include the detection of diseases in cucumber leaves and grapevines, as well as the discrimination of different grain types in heterogeneous granular mixtures [5–7]. Multispectral and hyperspectral imaging combine the strengths of conventional imaging and spectroscopy by capturing spatial and spectral information simultaneously, yielding a spectrum for each pixel. Multispectral imaging acquires a limited number of discrete wavelength bands, whereas hyperspectral imaging (HSI) captures tens to hundreds of contiguous spectral bands, providing a detailed spectral signature for every pixel. The data generated by HSI are commonly represented as a 3D data cube, or hypercube, consisting of two spatial dimensions and one spectral dimension. HSI systems frequently operate in the visible and near-infrared (VNIR, 400–1000 nm) and short-wave infrared (SWIR, 1000–2500 nm) regions of the electromagnetic spectrum. These spectral ranges contain information related to the chemical composition and physical properties of materials [8].

The high dimensionality and strong correlation between spectral bands present challenges for data analysis. Traditionally, HSI data have been analyzed using multivariate statistical methods such as principal component analysis (PCA), partial least squares regression (PLSR), soft independent modelling of class analogy (SIMCA), and partial least squares discriminant analysis (PLS-DA) [9]. These approaches primarily

exploit spectral information, treating pixels independently, and are therefore often referred to as 'spectral classifiers' [10]. More recently, advances in machine learning have been exploited to include spatial information in the analysis as well [11]. For this task, convolutional neural networks (CNNs) are a common choice [12].

HSI often combines rich spectral and spatial information, yet many tasks on publicly available HSI datasets can achieve strong performance using primarily either spectral or spatial features. For example, widely used remote sensing datasets such as Indian Pines and Kennedy Space Centre have been extensively studied using methods that rely predominantly on spectral signatures for classification [13]. Beyond remote sensing, HSI has also been widely applied in food science for the assessment of quality attributes [14], including moisture, fat, and protein content. An example is the CoaSpec dataset, which was developed for cocoa bean quality assessment [15]. However, many datasets in this domain consist of point-wise measurements using NIR spectroscopy rather than spatially resolved hyperspectral images [16]. Furthermore, the classification task in the relatively few available agrifood HSI datasets is focused on global sample properties, such as ripeness or quality attributes, for which mean spectra and conventional multivariate statistical methods already achieve high performance [17,18], as summarized in Table 1. At the same time, spectral-spatial classifiers have been developed for such applications. For instance, a recent study developed a spectral-spatial fusion network for detecting potato tuber dry rot, achieving an accuracy of about 98%, compared to 97% for a conventional spectral SVM classifier [19]. Likewise, in another study, spectral SVM classifiers achieved performance comparable to, or even exceeding, that of a 3D CNN for the classification of confectionary products and the discrimination between white stripe and red muscle tissue in salmon fillets [20]. For the classification of bruised plums, a spectral NPLS-DA classifier achieved an accuracy of 75%, compared with 87% for a spectral-spatial CNN. In contrast, a larger performance gap was observed for mango ripeness classification, where a novel spectral-spatial approach achieved an accuracy of 92% compared with 58% for a traditional spectral Random Forest classifier [21,22]. To the best of our knowledge, no publicly available dataset provides a benchmark specifically designed for developing and evaluating algorithms that must jointly exploit both spectral

and spatial information. The lack of such spectral–spatial benchmark datasets limits the development and evaluation of machine learning (ML) methods that fully leverage the information content of HSI data.

Table 1. Overview of publicly available hyperspectral imaging (HSI) datasets for classification of agrifood products. For each dataset, the number of full HSI images, camera system, spectral range, number of spectral bands, target classes, and class balance are listed. The final column indicates whether successful classification has been demonstrated using spectral features only (accuracy reported), thereby providing insight into the extent to which spatial information is required for the task. Spectral-only baseline results for our HyperCookies dataset are given in Table 2. NR, not reported.

Dataset	No. of samples	Camera	Range	No. of bands	Classes	Class balance	Successful spectral-only baseline?
HyperLeaf2024[23]	2410	SpecimIQ	400-1000 nm	204	4 cultivars (+ regression targets)	NR	NR
Canadian wild rice kernels[24]	2,048	Specim FX 10	400-1000 nm	224	8 quality classes	No	NR
Avocado[25]	180	Specim FX10	400-1000 nm	224	3 firmness levels	Yes	Yes (77.8%)
		INNO-SPEC Redeye	950-1700 nm	252	3 ripeness levels	Yes	Yes (66.7%)
Kiwi[25]	262	Specim FX10	400-1000 nm	224	3 firmness levels	Yes	Yes (60.9%)
		INNO-SPEC Redeye	950-1700 nm	252	3 sweetness levels		Yes (82.6%)
					3 ripeness levels		Yes (55.6%)
Grapevine leaves[26]	205	SpecimIQ	400-900 nm	204	healthy and 11 diseases	No	Yes (87.2%)[27]
Agricultural plant HSI[28]	385	Offner	420-979 nm	237	16 agricultural crops	No	Yes (96%)
HyperspectralBlueberries[29]	420	Pika XC2	400-1000 nm	462	2 (normal and defective)	Yes	Yes (95.7%)
Chicken eggs[30]	1,228	Pika L	374-1014 nm	300	2 (fertile and infertile)	Yes	Yes (90.3%)
HyperCookies (ours)	320	Specim FX10	400-1000 nm	224	4 topping patterns	Yes	No
		Specim FX17	900-1700 nm	224			

Therefore, the objective of this contribution is to develop, describe and make available HyperCookies, the first annotated HSI dataset specifically designed to require the integration of both spectral and spatial information for accurate classification. The dataset comprises four cookie classes that differ in both topping composition and topping arrangement. Spectral variation is introduced through chocolate droplets with 54% or 80% cocoa content, while spatial variation is represented by checkerboard and circular topping patterns. As a result, the classes cannot be reliably discriminated using neither spectral information alone, since the mean spectral signatures of the classes largely overlap, nor by spatial information alone, since the two chocolate types are visually indistinguishable. Consequently, accurate classification requires the joint exploitation of spectral and spatial information. To provide opportunities to investigate transfer learning, domain adaptation, and domain shift, HSI data were acquired in the VNIR and SWIR ranges, both before and after baking the cookies. In addition, predefined training, validation, and test splits are provided to facilitate reproducible benchmarking. Baseline results obtained using seven spectral and six image-based classifiers demonstrate the limitations of relying exclusively on either spectral or spatial information for such HSI datasets. These experiments should not be interpreted as a comprehensive benchmark of HSI classification methods. Rather, they were designed to illustrate the challenges associated with only spectral and spatial approaches and to highlight the need for methods that effectively integrate spectral and spatial information for HSI classification.

Methods

Sample description

The dataset comprises 160 handmade cookies prepared from a dough containing 60% w/w peanut butter (85.5% peanuts, sunflower oil, fully hydrogenated palm fat, salt), 26% w/w wheat flour, and 14% w/w powdered sugar, adapted from the recipe of Lane [31]. Ingredients were weighed using a laboratory balance (Kern PCB 10000-1, Kern & Sohn, Balingen, Germany, max 10,000 g, d = 0.1 g) according to fixed target proportions. The ingredients were mixed until a homogeneous dough was obtained. Two independent dough batches were prepared on different days, each yielding 80 cookies. Although

identical formulations and preparation procedures were used, the use of independent production batches introduces realistic batch-to-batch variability, reducing the risk of overly optimistic classifier performance that may arise when all samples originate from a single homogeneous batch. The dough was flattened to a thickness of approximately 0.8 cm using a flat plate and circular cookie bases were cut using an 8 cm diameter cookie cutter. Each cookie was decorated with chocolate droplets of two chocolate formulations: 54% and 80% cocoa chocolate. According to the manufacturer, the 54% chocolate contained cocoa mass, sugar, cocoa butter, soy lecithin, and natural vanilla flavour, whereas the 80% chocolate contained cocoa mass, sugar, low-fat cocoa powder, soy lecithin, and natural vanilla flavour. The two formulations therefore differed primarily in their cocoa content and the presence of added cocoa butter in the 54% chocolate. Both chocolates appeared visually similar in RGB images, making them suitable for HSI-based classification. Two spatial arrangements of chocolate droplets were used: circular and checkerboard patterns. For each pattern, the positions of the two chocolate types were alternated, resulting in four experimental classes, each consisting of 40 cookies, as shown in Figure 1. Chocolate droplets were placed manually, maintaining the intended pattern while allowing natural variation in the number, density, and exact position of droplets. Additional variability was introduced through repeated handling of the cookies, causing minor deformations from the ideal circular shape (Fig. 1). HSI data were acquired for all cookies in both unbaked and baked conditions, allowing the evaluation of domain shifts induced by physicochemical changes during baking. Following the initial image acquisition, the cookies were baked in an oven (Vötsch VTU 100/150) set at 180 °C for 12 minutes and allowed to cool and rest for approximately 24 hours prior to spectral imaging. A second batch of cookies was prepared and baked one day later. Due to practical constraints, two cookies from Class 1 and two from Class 4 were missing during acquisition of the second batch, resulting in 38 spectral hypercubes for Classes 1 and 4 and 40 for Classes 2 and 3.

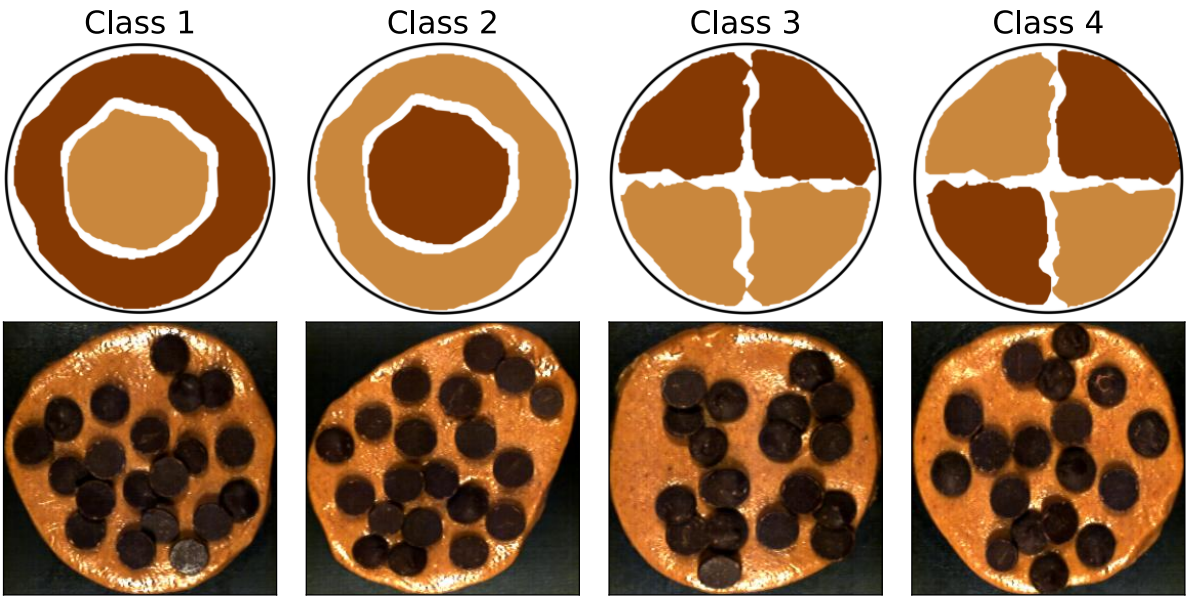


Figure 1. Schematic representation (top row) and RGB image (bottom row) of the four cookie classes. Class 1: circular pattern with 80% cocoa droplets in the outer ring and 54% cocoa droplets in the middle. Class 2: circular pattern with 54% cocoa droplets in the outer ring and 80% cocoa droplets in the middle. Class 3: checkerboard pattern with 80% cocoa droplets in the upper half and 54% cocoa droplets in the lower half. Class 4: checkerboard pattern with 80% and 54% cocoa droplets arranged diagonally.

Image acquisition

The image acquisition setup consisted of a moving sample bed and two hyperspectral pushbroom cameras covering the VNIR (Specim FX10, 400-1000 nm) and SWIR range (Specim FX17, 900-1700 nm), as shown in Figure 2. The cameras are line-scan cameras that scan the object line by line to obtain a hyperspectral image cube with two spatial dimensions and a spectral dimension. The FX10 camera simultaneously acquires 224 spectral bands between 398.75 and 1002 nm, with a sampling interval of ~2.56 nm and a spectral resolution (FWHM) of ~2.5–2.6 nm. The FX17 camera simultaneously acquires 224 bands between 935.6 and 1720 nm, with a sampling interval of ~3.46 nm and a FWHM of ~3.5 nm. The samples were illuminated by two halogen lamps (Osram Light AG, 50 W, 12 V DC) at each camera, with one lamp positioned on either side of the moving sample bed. After the lamps were switched on, they were allowed to warm up for 30 min to ensure stable illumination.

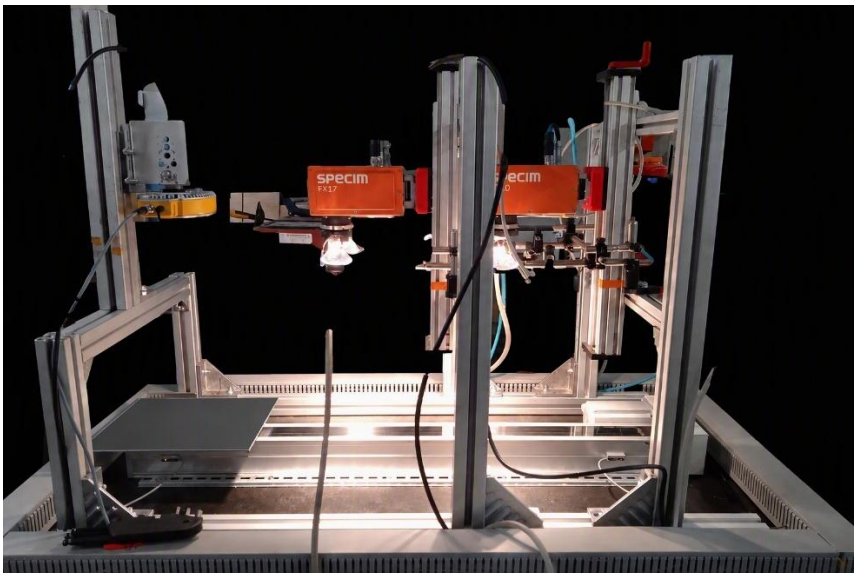


Figure 2. Hyperspectral image acquisition system consisting of three cameras: an RGB camera (left), a SWIR camera (SPECIM FX17; middle), and a VNIR camera (SPECIM FX10; right). The blue plate is the sample tray, which moves beneath the cameras during image acquisition. Original image adapted through GenAI for removing the sample on the tray.

To convert raw images into reflectance data, white and dark reference images were acquired for each camera. The white reference was obtained using a Spectralon calibration panel, while the dark reference was acquired with the camera shutter closed. These references were used to correct for non-uniform illumination, sensor sensitivity variations, electronic noise, and thermal effects according to the following equation:

$$R = \frac{I - D}{W - D} \quad (1)$$

where R is the calibrated reflectance image, I is the raw image, W is the white reference image, and D is the dark reference image.

Training, validation, and test data

The dataset comprised 160 cookies collected over two acquisition days, with 80 cookies (20 samples per class) measured on each day. To account for day-to-day variability, the samples from each acquisition day were randomly divided into two equal subsets while maintaining the original class balance, resulting in four stratified subsets in total. For model training, two subsets were assigned to the training set (50% of the data), one subset to the validation set (25%), and one subset to the test set (25%). Considering all

possible assignments of the four subsets resulted in twelve unique train-validation-test combinations.

Data preprocessing and calibration

Each hyperspectral acquisition contained two cookies positioned on the tray that is shown in Figure 2. The acquired raw hypercube, consisting of the line-scan frames together with the corresponding dark and white reference scans, was converted into a calibrated reflectance cube following Eq. (1). Any negative reflectance values, which occurred mainly in the edge bands of the FX10, were clipped to zero. The cookies in each scene were then located by computing the per-pixel spectral variance, defined as the variance of the reflectance values across all wavelength bands for each pixel individually. This variance is close to zero on the flat dark tray and high on the textured cookie surface. The spectral variance image was then thresholded using Otsu's method [32]. The resulting mask was used only to determine the bounding box of each cookie, so the crop retained the real background pixels instead of setting them to zero. Each cropped cube was then bilinearly resized along the two spatial axes to 256 × 256 pixels. An overview of the complete preprocessing and calibration workflow is provided in Figure 3.

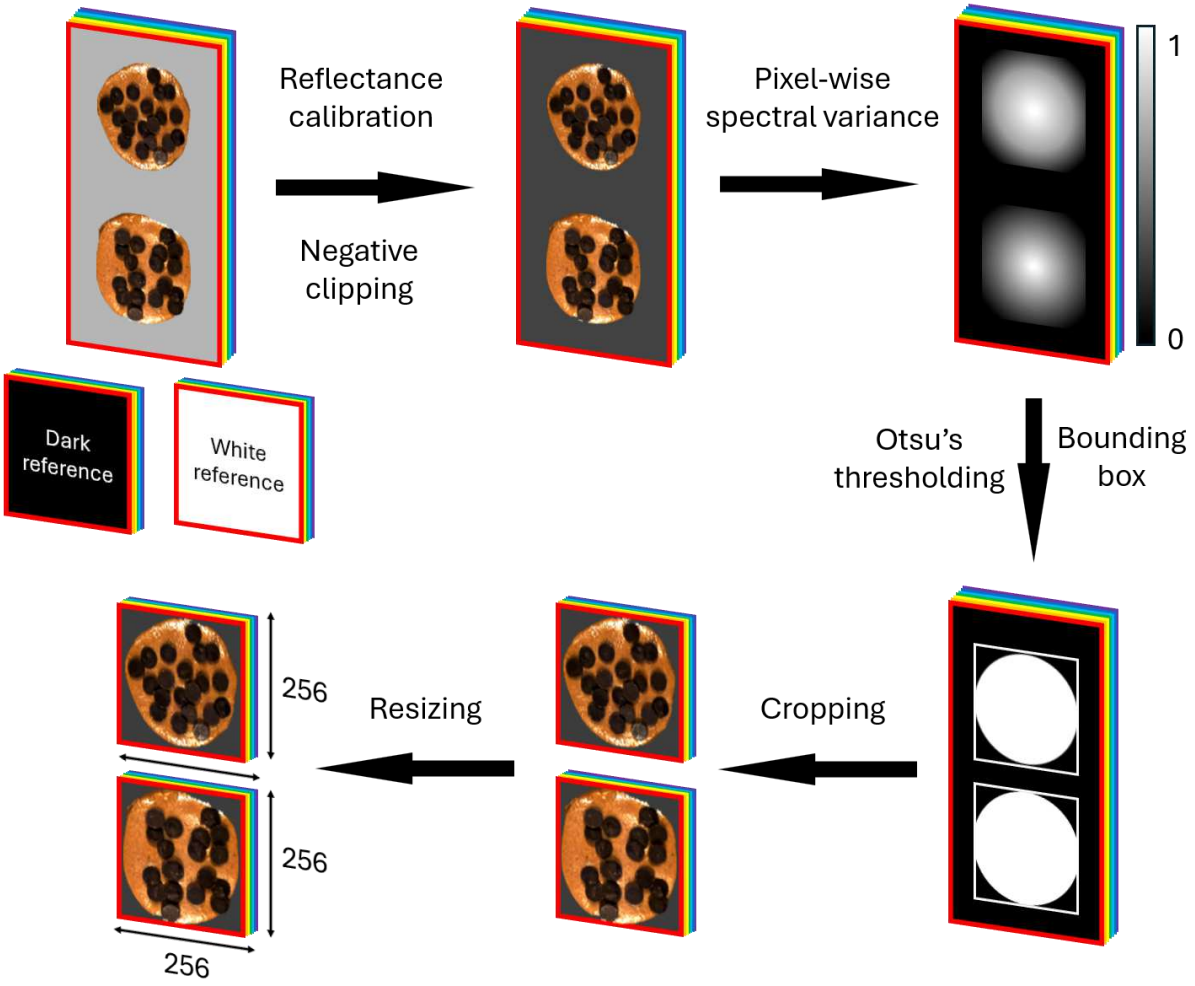


Figure 3. Data preprocessing and calibration workflow. The acquired intensity hypercubes were calibrated using dark and white reference scans, followed by clipping of negative reflectance values. Cookies were identified by applying Otsu thresholding to the logarithm of the pixel-wise spectral variance, and the resulting bounding boxes were used to crop the reflectance cubes. The cropped cubes were subsequently resized to 256 × 256 pixels along the spatial dimensions.

Classification using spectral and image-based classifiers

The hyperspectral data were classified using either purely spectral or purely spatial (image-based) classifiers. These experiments should not be interpreted as a comprehensive benchmark of HSI classification methods. Rather, they were selected to highlight the challenges associated with relying exclusively on spectral or spatial information and to demonstrate the importance of jointly exploiting both spectral and spatial features for accurate hyperspectral image classification.

Input representations

Three input representations were considered. (i) Pseudo-RGB images, used only as inputs to the image-based classifiers, were generated for the VNIR data by selecting the bands at 638.9 nm, 550.2 nm, and 460.5 nm, corresponding to the red, green, and blue wavelengths. (ii) Three-channel PCA composites were generated for both the VNIR and SWIR data from the score images of the first three principal components. (iii) The full VNIR and SWIR hyperspectral cubes were used without reduction.

Spectral classifiers

The spectral classifiers do not exploit spatial information. The full hyperspectral cubes and the PCA composites were averaged over the spatial dimensions, yielding a single vector with the mean reflectance for each spectral band or the mean score for each principal component. These vectors were subsequently used as input to conventional ML classifiers, including Gaussian Naïve Bayes (GNB) [33], k-Nearest Neighbors (kNN) [34], Linear Discriminant Analysis (LDA) [35], Logistic Regression (LR) [36], Partial Least Squares Discriminant Analysis (PLS-DA) [9], Random Forest (RF) [37], and Support Vector Machine (SVM) [38], implemented using the Scikit-Learn library [39]. Hyperparameter optimization was performed using a grid search strategy on the validation set.

Image-based classifiers

The pseudo-RGB images, three-channel PCA composites, and the full VNIR and SWIR hyperspectral cubes were used as input data. For the latter, the original hyperspectral cubes were reduced to three channels through a dimensionality reduction module consisting of a single convolutional layer with three output channels, which was placed before the CNN backbone. In another case, the first pretrained layers of the network were replaced by a newly initialized layer whose number of input channels matched the spectral dimension of the hyperspectral cube.

Six conventional CNN architectures, namely ResNet18, ResNet50 [4], EfficientNet-V2-S, EfficientNet-V2-M [40], MobileNet-V3-S, and MobileNet-V3-L [41] from the Torchvision library, were trained for each input configuration. Given the limited size of the training dataset, ImageNet-pretrained weights were used as a form of regularization. The networks were trained using cross-entropy loss and the Adam optimizer with a weight decay of 10^{-3}. An

initial learning rate of 10^{-3} was used and reduced by a factor of 10 whenever the validation loss reached a 3-epoch plateau. Training was performed for 100 epochs, and the model weights corresponding to the lowest validation loss were retained for inference on the test set.

Evaluation metrics

The classification algorithms were evaluated on twelve train-validation-test combinations. The predicted labels were compared with the ground-truth labels of the four classes (Class 1 to Class 4). Classification performance was quantified using accuracy, defined as the proportion of correctly classified samples relative to the total number of samples. Due to practical constraints, two hyperspectral hypercubes were missing from Class 1 and Class 4 in the baked cookie dataset, resulting in a slight class imbalance. Consequently, balanced accuracy was reported for the baked cookie experiments, whereas standard accuracy was reported for the perfectly balanced unbaked-cookie dataset.

Data Records

File formats and folder structure

The dataset is organized into two top-level categories, unbaked and baked, each following an identical directory structure. Each category is further divided into two subsets corresponding to the measurement day. Within each subset, data are organized according to the acquisition camera: FX10 (VNIR) and FX17 (SWIR). For each camera, one subdirectory is provided: hsi/ (hyperspectral image cubes stored in HDF5 (.h5) format).

Each sample follows the naming convention:
`<category>_<camera>_<day>_class<N>_<UID>`

where:

- `<category>` ∈ {`unbaked, baked`}
- `<camera>` ∈ {`FX10, FX17`}
- `<day>` ∈ {`Day1, Day2`}
- `<N>` denotes the class label
- `<UID>` is a unique sample identifier.

The complete directory structure is shown below:

```
Data/
|-unbaked/
|   |-Day1/
|   |   |-FX10/
|   |   |   |-hsi/
|   |   |-FX17/
|   |   |   |-hsi/
|   |-Day2/ [same structure as Day1/]
|-baked/ [same structure as unbaked/]
```

HSI cube files

Each HDF5 file contains the data acquired for a single cookie stored as a 224 × 256 × 256 float32 reflectance cube (spectral bands × height × width). The wavelengths and per-band FWHM are stored as attributes (wavelength_nm, fwhm_nm), together with metadata describing the bake state, camera, class, split, source file, and processing steps applied.

Additional metadata

In addition to the class labels, the dataset includes a metadata file (`info_observations.csv`) describing sample-specific observations. For the baked cookie samples, instances exhibiting smeared chocolate droplets or surface cracks were recorded. These annotations enable researchers to identify and optionally exclude such samples, or to investigate the impact of baking-related features on classification performance and model robustness.

Dataloader

The released code includes two PyTorch Dataset objects for loading in the HSI and RGB files, respectively. The object has camera, baked, fold, and subset as its parameters. The `info_unbaked.csv` and `info_baked.csv` are included for reproducing the train, validation, and test splits proposed in this work.

Technical Validation

The technical validation of the HyperCookies dataset includes a spectral quality analysis demonstrating the discriminative power of the two chocolate drop types (54% vs 80% cocoa) and the importance of incorporating spatial information in addition to spectral information for the classification of the four cookie classes. Furthermore, baseline experiments are included to assess the performance of spectral and image-based (spatial) classifiers trained on the HSI data.

Visual assessment

Figure 4 provides a visual assessment of the classification task. Before baking, the four classes are visually indistinguishable, as both chip types appear nearly identical against the raw dough in the pseudo-RGB and false-color images. After baking, differences between the two chip types become partially discernible, not only through subtle color variations but also through changes in surface texture. This distinction is more pronounced in the FX17 false-color composites than in the FX10 images. In addition, baking introduced some visual artifacts, as several chips smeared into the surrounding dough and some cookies developed slight cracks, partially obscuring the original chip arrangements.

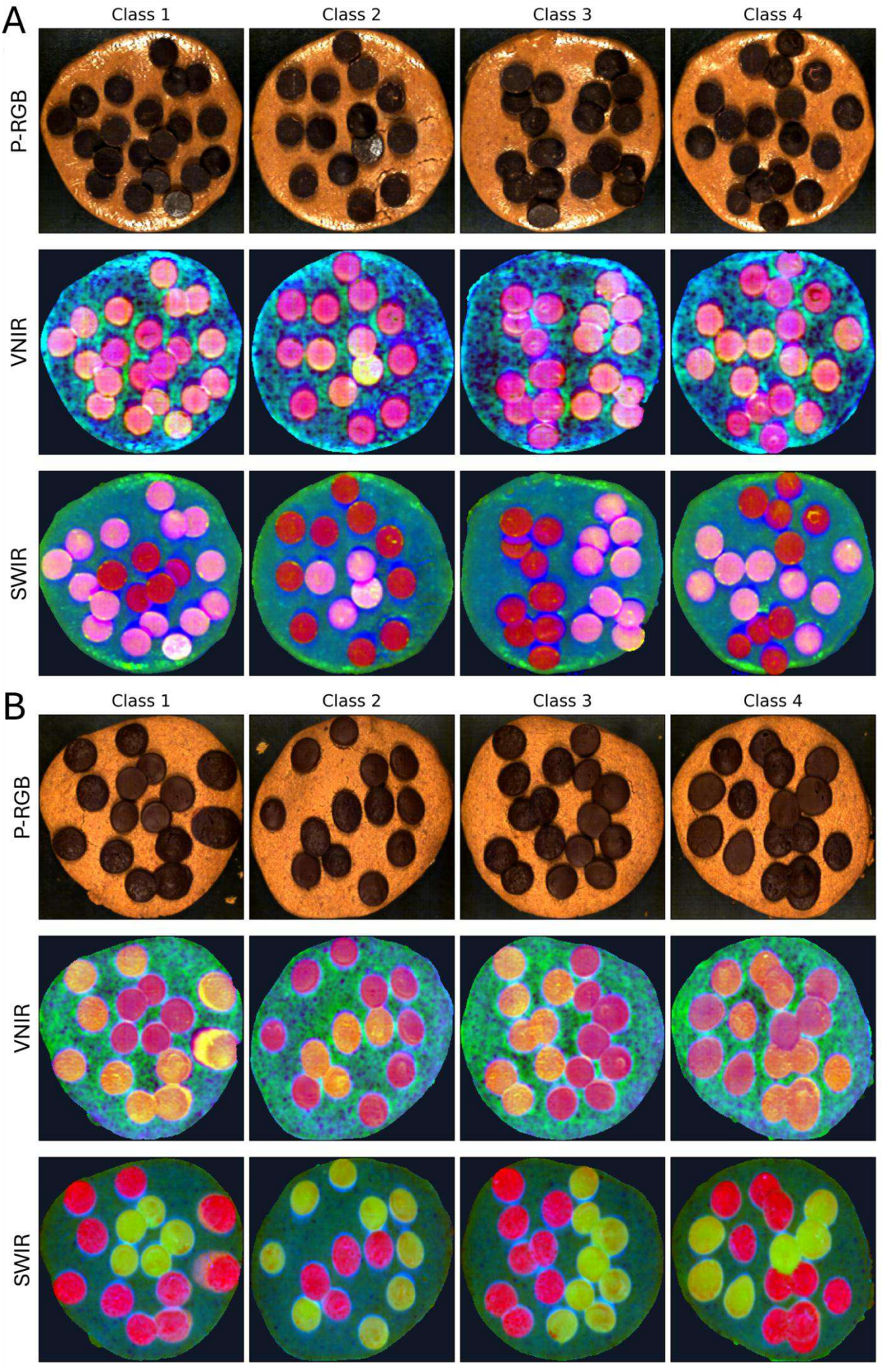
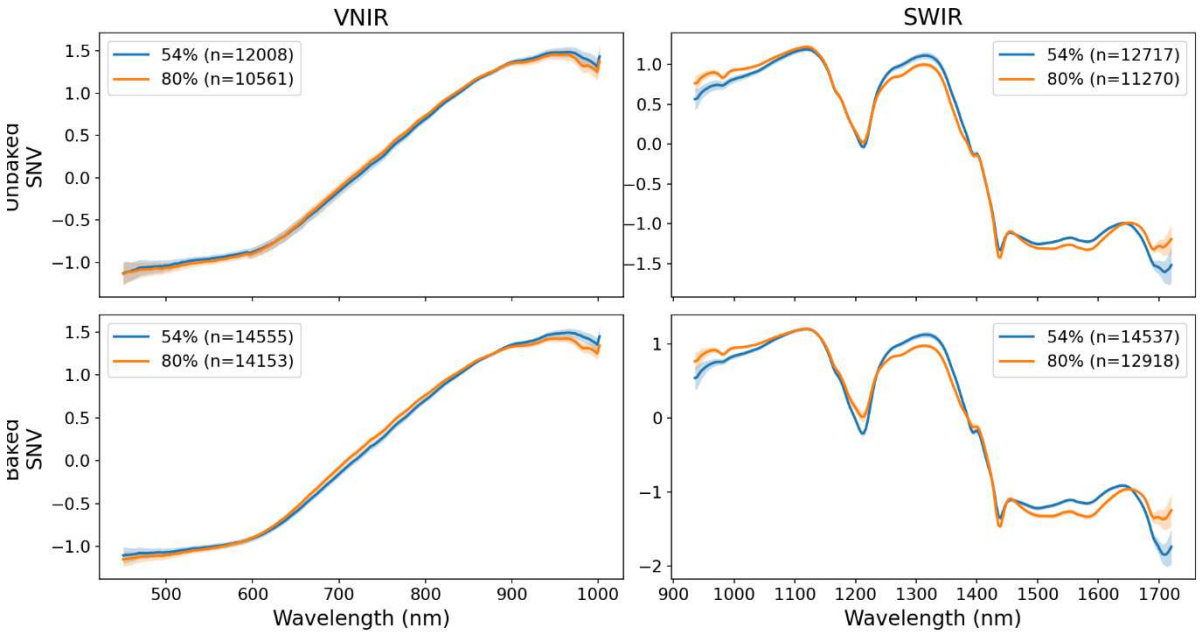

A
Class 1
Class 2
Class 3
Class 4
P-RGB
VNIR
SWIR
B
Class 1
Class 2
Class 3
Class 4
P-RGB
VNIR
SWIR

Figure 4. Representative samples from each cookie class in (A) the unbaked and (B) the baked state. Pseudo-RGB (P-RGB) representations generated from bands at 638.9, 550.2, and 460.5 nm, and composites generated from the first three principal components for the VNIR and SWIR data.

Spectral Quality Analysis

The spectral quality of the samples was evaluated using mean spectral signatures extracted from HSI data of 54% and 80% cocoa droplets. To obtain these signatures, one representative cookie from each class was manually annotated, using the 'labelme' python package [42]. Ground-truth droplet labels were assigned to the corresponding chocolate formulation based on visual inspection of the PC score images. Two cookies from Class 1 and Class 2 were analyzed per day, and three droplets for each chocolate formulation were segmented in every cookie. The resulting droplet masks were applied to the original HSI cubes to extract the spectra of each chocolate type, from which the mean spectral signatures and standard normal variate preprocessed ones were calculated [43]. The VNIR data (FX10) shown in Figure 5 indicate that, for both the unbaked and baked versions, the 80% cocoa droplets have a slightly higher reflectance than the 54% cocoa droplets between 650 and 900 nm. The spectral signatures obtained with the SWIR camera (FX17) are presented in the right part of Figure 5. In the SWIR range, the 80% cocoa droplets also show higher reflectance up to 1250 nm. Overall, the spectra show expected absorption dips associated with C-H bonds present in lipids and O-H bonds in sucrose around 900 nm, 1200 nm, and 1400 nm, consistent with earlier work [15,44].



354

355 *Figure 5. Mean spectral signatures after standard normal variate preprocessing of the two*
356 *cocoa chocolate droplets (54% and 80%) in the unbaked and baked states, obtained in*
357 *the VNIR (FX10) and SWIR (FX17) spectral ranges. The shaded region around each mean*
358 *spectrum represents ± 1 standard deviation calculated over the pixels. 'n' indicates the*
359 *number of pixels over which the mean and standard deviation was calculated.*

360 A second spectral analysis was conducted on the mean spectra of the four cookie
361 classes, considering both unbaked and baked samples. Figure 6 demonstrates that the
362 classes cannot be reliably discriminated based solely on their spectral signatures, either
363 in the VNIR (FX10) or SWIR (FX17) ranges. This result is not unexpected, since averaging
364 the spectra removes spatial information related to the positioning and distribution of the
365 two chocolate droplet types [45]. Consequently, the spatial arrangement of the droplets
366 must be taken into account to achieve accurate classification.

367

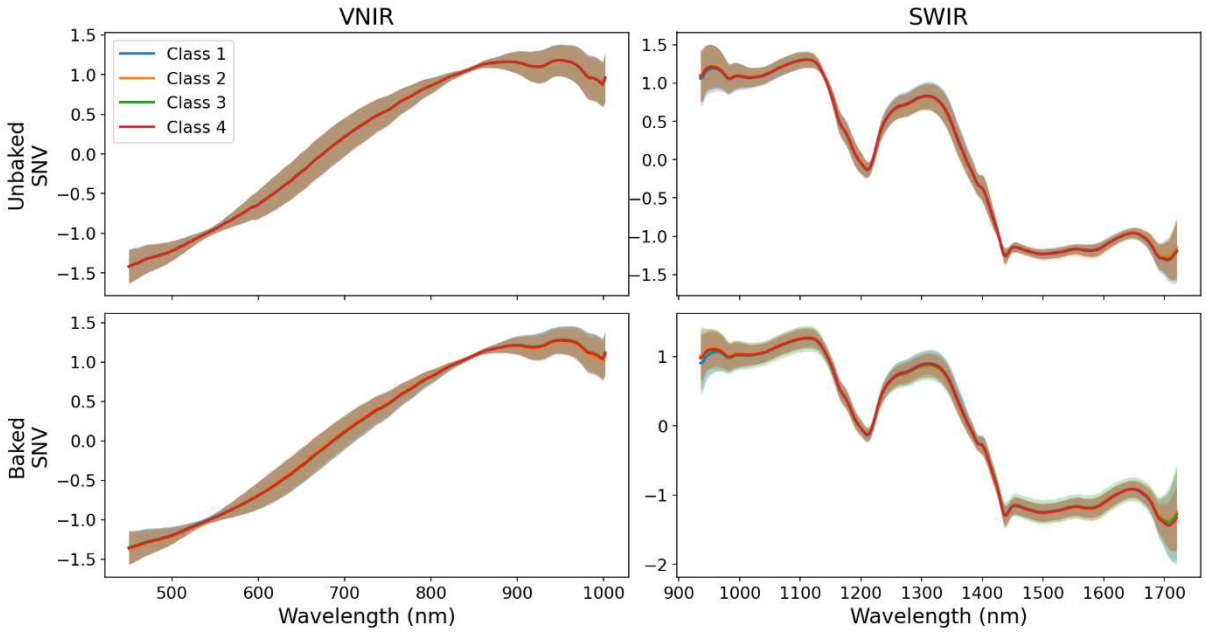


Figure 6. Mean spectral signatures after standard normal variate preprocessing of the four cookie classes in the unbaked and baked states, obtained in the VNIR (FX10) and SWIR (FX17) spectral ranges. The shaded region around each mean spectrum represents ± 1 standard deviation calculated over the number of cookies per class.

Classification Baseline

The experiments presented in this section are intended to illustrate the difficulty of classifying the data using either purely spectral or purely image-based (spatial) classifiers.

Spectral classifiers

Seven spectral classifiers were trained and evaluated using image-level mean spectra obtained by averaging the spectral information across all spatial pixels. The models were trained using either the full hyperspectral data or the mean scores on the first three principal components (PCs). The results presented in Table 2 for the unbaked and baked samples indicate that the classifiers were unable to reliably distinguish between the four predefined classes. On the VNIR (FX10) data, all models for the unbaked samples achieved accuracies between 20% and 30%, corresponding to performance at or only slightly above random guessing (25%), regardless of whether the full spectra or the three PCs were used. On the SWIR (FX17) data, accuracies for the unbaked samples were consistently higher (31-46%), in line with the stronger chemical contrast between the two drop types in this spectral range (Figure 5). For the baked samples, the models achieved

accuracies between 28% and 43% on the VNIR data, and between 32% and 53% on the SWIR data. Overall, the performance on the baked samples was slightly higher than on the unbaked samples. Since these approaches rely solely on spectral information and do not exploit spatial context, the limited performance is, however, not unexpected. The accuracy slightly above chance level may be explained by differences in the overall spectral characteristics of the classes. For example, Classes 1 and 2 differ only in the placement of the 80% and 54% cocoa droplets within the circular pattern. This can result in different surface coverage and numbers of droplets for each chocolate type, possibly introducing subtle spectral differences that provide limited discriminative information. Reducing the spectra to their first three PCs did not lead to a systematic improvement of the best performing models, suggesting that the subtle class-related spectral differences are not captured fully by the directions of largest spectral variance. Overall, none of the classifiers were able to reliably distinguish the classes.

Table 2. Classification accuracy (%) on the test set for the different spectral classifiers using the first three principal components (PC) or the full HSI data as input. Values are mean ± standard deviation over the twelve splits. GNB, Gaussian Naïve Bayes; kNN, k-Nearest Neighbors; LDA, Linear Discriminant Analysis; LR, Logistic Regression; PLS-DA, Partial Least Squares Discriminant Analysis; RF, Random Forest; SVM, Support Vector Machine.

	VNIR				SWIR			
	Unbaked		Baked		Unbaked		Baked	
	PC	HSI	PC	HSI	PC	HSI	PC	HSI
GNB	29.4±6.1	24.6±7.0	30.9±7.0	28.3±7.4	36.9±7.3	35.0±5.2	36.7±14.2	31.9±7.9
kNN	27.1±6.4	21.2±9.6	36.7±11.6	28.5±5.1	34.2±7.3	32.9±8.4	39.7±13.1	35.4±7.6
LDA	23.1±5.7	29.6±7.9	27.8±7.3	31.6±10.0	35.0±6.7	**46.0±14.2**	45.0±8.9	**53.0±13.5**
LR	20.6±4.1	30.0±7.9	26.6±5.3	30.0±7.3	32.5±6.0	43.3±9.8	34.3±9.8	45.8±13.2
PLS-DA	22.3±4.6	**30.4±6.1**	29.1±6.3	**42.7±12.5**	34.6±5.2	40.6±12.3	41.9±10.4	52.7±9.5
RF	26.5±4.3	24.8±6.3	36.1±8.1	31.9±8.0	30.8±6.9	34.4±6.8	39.1±12.9	34.3±8.2
SVM	26.0±5.7	25.6±6.2	32.5±9.9	30.0±10.6	35.4±8.4	40.4±8.8	45.5±11.3	39.8±7.2

Image-based classifiers

Six image-based classifiers (CNNs) were trained and evaluated on the unbaked and baked cookies. To match the spectral dimensions for the HSI data and the three-channel input expected by standard CNN architectures, four approaches were explored: pseudo-RGB representations for the VNIR data, composites generated from the first three principal components, dimensionality reduction, and the addition of an extra convolutional layer. The results presented in Table 3 indicate that the classifiers were unable to reliably distinguish between the four predefined classes. The high standard deviations in classification accuracy across the twelve folds further suggest inconsistent model convergence. This was evident from the unstable training and validation loss curves, which failed to reach a clear plateau. Together, these findings indicate that the models struggled with the classification task and that spatial information alone was insufficient to reliably discriminate between classes. Overall, models trained on SWIR data slightly outperformed those trained on VNIR data, with the three-channel principal component representation of the SWIR data achieving the highest overall performance of 43.8% on the unbaked cookies. While the spectral classifiers performed slightly better on the baked samples, the image-based classifiers in Table 3 did not noticeably perform better for the baked than for the unbaked samples.

Table 3. Classification accuracy (%) obtained with different CNNs on the VNIR (FX10) and SWIR (FX17) test sets using four approaches for converting the hyperspectral input into a three-channel representation: pseudo-RGB (P-RGB) (only for VNIR data), composite from the first three principal components (PC), input dimensionality reduction (DR), and an additional convolutional layer (Conv). Values are reported as mean ± standard deviation. EffNet, EfficientNet-V2; MobNet, MobileNet-V3; S, small; M, medium; L, large.

	VNIR								SWIR					
	Unbaked				Baked				Unbaked			Baked		
	P-RGB	PC	DR	Conv	P-RGB	PC	DR	Conv	PC	DR	Conv	PC	DR	Conv
ResNet18	24.6±7.7	25.0±4.9	24.6±5.8	26.7±4.2	**37.8±11.1**	33.6±6.7	36.1±7.7	29.0±7.4	**43.8±9.1**	28.1±6.7	35.8±13.0	41.1±11.5	36.1±9.3	41.9±13.9
ResNet50	25.4±5.6	28.3±6.7	30.6±7.4	25.6±3.7	32.7±7.0	32.7±8.2	38.9±5.5	24.7±6.3	31.2±5.3	31.0±5.1	36.7±9.0	36.8±9.9	36.4±9.1	**43.1±13.5**
EffNet-S	29.8±9.9	**34.4±6.8**	29.2±6.1	23.3±4.7	26.3±9.7	28.4±8.0	28.7±7.3	24.4±4.8	37.1±9.3	26.0±7.6	24.2±7.3	28.4±7.2	31.1±9.0	27.0±8.3
EffNet-M	24.6±5.5	26.9±6.2	27.3±5.1	29.2±3.7	26.2±8.5	28.9±5.5	27.8±4.8	27.3±6.8	30.2±6.9	25.2±8.1	28.8±8.5	31.4±10.5	32.0±8.5	25.6±9.9
MobNet-S	28.1±4.3	24.0±7.9	26.7±5.5	25.6±2.4	35.9±8.0	31.2±8.3	32.3±9.7	27.7±4.7	35.2±8.9	27.7±4.7	24.8±2.9	34.9±15.9	28.2±6.9	27.2±7.5
MobNet-L	28.3±4.7	24.4±4.5	24.0±1.7	24.2±3.3	32.4±9.5	32.8±8.4	35.6±7.0	23.5±4.1	31.0±6.5	25.8±5.5	24.2±2.7	36.4±14.1	32.7±6.3	24.4±3.5

In summary, none of the evaluated classifiers achieved satisfactory and consistent classification performance. While SWIR data yielded slightly better results than VNIR data, performance remained well below the level required for reliable discrimination between classes. Overall, these findings highlight the need for more advanced approaches that can jointly exploit spectral and spatial information to address the complexity of the classification task. As such, the dataset provides a valuable benchmark for the development of spatial-spectral ML models.

Usage Notes

The HyperCookies dataset was designed as a hyperspectral benchmark dataset, encouraging future ML models to rely on both spatial and spectral information for classification. It contains cookies with visually similar 54% and 80% cocoa droplets arranged in different patterns and presented in both unbaked and baked states. This requires models to jointly identify chocolate droplet composition and spatial arrangement of the droplets. The dataset is intended primarily for benchmarking robust spatial-spectral classification, sensor-specific modelling, domain-shift handling, and generalization under realistic physical variability.

The SWIR subset (900–1700 nm) provides stronger chemical discrimination through absorption features associated with cocoa butter and sucrose, as demonstrated in Figure 5. It is particularly suitable for studying spatial feature extraction after spectral separation. In contrast, the VNIR subset (400–1000 nm) exhibits weaker spectral contrast and greater sensitivity to illumination drift, making it well suited for evaluating drift-resistant preprocessing and classification methods. Therefore, results obtained from VNIR and SWIR data should be reported separately, with the sensor, preprocessing, partitioning strategy, and handling of batch effects clearly documented. In addition, researchers may evaluate the added value of multimodal spectral fusion by comparing models trained on VNIR data, SWIR data, and their combination.

Dataset variability was deliberately introduced through the random manual placement of chocolate droplets and the use of two independent measurement days. To support a fair and robust evaluation, the dataset is provided with twelve predefined train-validation-test splits (CSV files), stratified by class to ensure a balanced distribution of samples

across the training, validation, and test sets. Users are encouraged to evaluate models across all predefined splits and report the mean and standard deviation of the selected metrics. This reduces the influence of any individual data partition and enables consistent comparison between models and analysis methods.

Data Availability

The HyperCookies dataset is publicly available at https://rdr.kuleuven.be/previewurl.xhtml?token=c0e19bcc-1b11-4cef-9f9f-a3d9adb90c8c (temporary link will become public upon paper acceptance). The dataset is released under a Creative Commons Attribution NoDerivatives 4.0 International (CC-BY-ND-4.0) license.

Code Availability

The supplementary code is publicly available at https://gitlab.kuleuven.be/u0168957/HyperCookies. The results in the paper can be reproduced by running scripts/benchmark_non_deep.sh and scripts/benchmark_deep.sh, after updating the root folder location in each script. The Figures can be reproduced by running process_results/plotting/create_figures.ipynb.

Acknowledgements

The authors would like to express their sincere gratitude to undergraduate students Demi Fatingan, Rik Maes, Anders Verbeeck, and Rhüne Vranken who assisted in the lab work.

Author Contributions

Conceptualization: all, Methodology and Data Acquisition: GM, AT, Software: GM, MS, Visualization: GM, NS, Data Curation: GM, Writing – Original Draft: AT, Writing – Review & Editing: all, Funding Acquisition: WS.

Funding

The authors would like to express their sincere gratitude to all collaborators in the SPECTR.AI project (Spectral image Processing with Efficiently Compressed TensoRs and AI, KU Leuven Industrial Research Fund, C2E/23/007).

Competing Interests

The authors declare no competing interests.

References

1. Mehta, N., Shaik, S., Devireddy, R. & Gartia, M. R. Single-Cell Analysis Using Hyperspectral Imaging Modalities. *Journal of Biomechanical Engineering* vol. 140 Preprint at https://doi.org/10.1115/1.4038638 (2018).

2. Deng, J. *et al.* ImageNet: A large-scale hierarchical image database. 248–255 (2009) doi:10.1109/cvprw.2009.5206848.

3. Krizhevsky, A. *et al.* ImageNet Classification with Deep Convolutional Neural Networks. *Proceedings of the 31st International Conference on Machine Learning (ICML-14)* **3**, (2012).

4. He, K., Zhang, X., Ren Shaoqing & Sun, J. Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (2016) doi:10.1002/chin.200650130.

5. Puyenbroeck, E. Van & Saeys, W. Cost-efficient training of hyperspectral deep learning models for the detection of contaminating grains in bulk oats by fluorescent tagging. *Spectrochim. Acta A Mol. Biomol. Spectrosc.* **332**, (2025).

6. Saffari, M., Mireei, S. A., Massah, A. & Saeys, W. Deep learning-enhanced Vis–NIR spectroscopy for early and stage-specific detection of cucumber mosaic virus in cucumber leaves with explainable models. *Biosyst. Eng.* **271**, (2026).

7. Mora Jimena, G., De Ketelaere, B. & Saeys, W. Tensor-based hyperspectral classification for grapevine disease detection: overcoming heterogeneous symptom distribution with limited training data. in *Photonic Technologies in Plant and Agricultural Science III* (eds. Heinemann, D. & Polder, G.) 8 (SPIE, 2026). doi:10.1117/12.3080503.

8. Nicolaï, B. M. *et al.* Nondestructive Measurement of Fruit and Vegetable Quality. *Annu. Rev. Food Sci. Technol.* **5**, 285–312 (2014).

9. Barker, M. & Rayens, W. Partial least squares for discrimination. *J. Chemom.* **17**, 166–173 (2003).

10. López-Maestresalas, A. *et al.* Non-destructive detection of blackspot in potatoes by Vis-NIR and SWIR hyperspectral imaging. *Food Control* **70**, 229–241 (2016).

11. Li, S. *et al.* Deep Learning for Hyperspectral Image Classification: An Overview. https://doi.org/10.1109/TGRS.2019.2907932 (2019) doi:10.1109/TGRS.2019.2907932.

12. Tempelaere, A. *et al.* An introduction to artificial intelligence in machine vision for postharvest detection of disorders in horticultural products. *Postharvest Biol. Technol.* **206**, 112576 (2023).

13. Zhong, Z., Li, J., Luo, Z. & Chapman, M. Spectral-Spatial Residual Network for Hyperspectral Image Classification: A 3-D Deep Learning Framework. *IEEE Transactions on Geoscience and Remote Sensing* **56**, 847–858 (2018).

14. Lu, Y., Saeys, W., Kim, M., Peng, Y. & Lu, R. Hyperspectral imaging technology for quality and safety evaluation of horticultural products: A review and celebration of the past 20-year progress. *Postharvest Biology and Technology* vol. 170 Preprint at https://doi.org/10.1016/j.postharvbio.2020.111318 (2020).

15. Contreras, K., Jouni, M., Dalla Mura, M. & Bacca, J. A Multimodal hyperspectral dataset of cocoa beans with physicochemical annotation. *Scientific Data* **12**, (2025).

16. Xing, F. & Chen, M. Advances in Hyperspectral Imaging for Nondestructive Food Quality and Safety Detection. *Foods* vol. 15 Preprint at https://doi.org/10.3390/foods15101631 (2026).

17. Varga, L. A., Makowski, J. & Zell, A. Measuring the Ripeness of Fruit with Hyperspectral Imaging and Deep Learning. in *Proceedings of the International Joint Conference on Neural Networks* vols 2021-July (Institute of Electrical and Electronics Engineers Inc., 2021).

18. Ryckewaert, M. *et al.* Hyperspectral images of grapevine leaves including healthy leaves and leaves with biotic and abiotic symptoms. *Sci. Data* **10**, (2023).

19. Cao, H. *et al.* Early Detection of Postharvest Potato Tuber Dry Rot Based on Hyperspectral Imaging and a Dual-Branch ResNet12-SE Spatial–Spectral Fusion Network. *Journal of Fungi* **12**, 702 (2026).

20. Zhu, H., Gowen, A., Feng, H., Yu, K. & Xu, J. L. Deep spectral-spatial features of near infrared hyperspectral images for pixel-wise classification of food products. *Sensors (Switzerland)* **20**, 1–20 (2020).

21. Castillo-Girones, S. *et al.* Detection of subsurface bruises in plums using spectral imaging and deep learning with wavelength selection. *Postharvest Biol. Technol.* **207**, (2024).

22. Jimena, G. M., De Ketelaere, B. & Saeys, W. Shared subspace learning via partial Tucker decomposition for hyperspectral image classification. *Spectrochim. Acta A Mol. Biomol. Spectrosc.* **343**, (2025).

23. Laprade, W. M. *et al.* HyperLeaf2024 - A Hyperspectral Imaging Dataset for Classification and Regression of Wheat Leaves. in *IEEE Computer Society Conference*

on Computer Vision and Pattern Recognition Workshops 1234–1243 (IEEE Computer Society, 2024). doi:10.1109/CVPRW63382.2024.00130.

24. Sikiru, Y. & Erkinbaev, C. A multimodal imaging dataset for quality grading of Canadian wild rice kernels using RGB and VNIR hyperspectral data. *Data Brief* **65**, 112601 (2026).

25. Varga, L. A., Makowski, J. & Zell, A. Measuring the Ripeness of Fruit with Hyperspectral Imaging and Deep Learning. in *Proceedings of the International Joint Conference on Neural Networks* vols 2021-July (2021).

26. Ryckewaert, M. *et al.* Hyperspectral images of grapevine leaves including healthy leaves and leaves with biotic and abiotic symptoms. *Sci. Data* **10**, (2023).

27. Meena, J. & Narayanan Unni, S. Effective Multiclass Grapevine Leaves Disease Classification Using Hyperspectral Imaging and Optimized Artificial Neural Network. *Cureus Journal of Computer Science* https://doi.org/10.7759/s44389-025-05883-5 (2025) doi:10.7759/s44389-025-05883-5.

28. Gaidel, A. V. *et al.* Agricultural plant hyperspectral imaging dataset. *Computer Optics* **47**, 442–450 (2023).

29. Deng, B., Lu, Y. & Stafne, E. Fusing spectral and spatial features of hyperspectral reflectance imagery for differentiating between normal and defective blueberries. *Smart Agricultural Technology* **8**, 100473 (2024).

30. Ahmed, M. W., Song, D., Ahmed, M. T. & Kamruzzaman, M. Hyperspectral imaging dataset for non-destructive fertility and structural evaluation of chicken eggs. *Scientific Data* **13**, (2026).

31. Lane, C. 2 Ingredient Peanut Butter Cookies. https://www.dessertfortwo.com/2-ingredient-peanut-butter-cookies/ (2022).

32. Otsu, N. *et al.* A Threshold Selection Method from Gray-Level Histograms. *IEEE Trans. Syst. Man Cybern.* **C**, 62–66 (1979).

33. John, G. H. *Estimating Continuous Distributions in Bayesian Classifiers*. http://robotics.stanford.edu/-gjohn/ (1955).

34. Cover, T. M. & Hart, P. E. *Nearest Neighbor Pattern Classification*. *IEEE TRANSACTIONS ON INFORMATION THEORY* vol. 24 (1952).

35. Fisher, R. A. The Use of Multiple Measurements in Taxonomic Problems. *Ann. Eugen.* **7**, 179–188 (1936).

36. Cox, D. R. *The Regression Analysis of Binary Sequences*. *Source: Journal of the Royal Statistical Society. Series B (Methodological)* vol. 20 https://about.jstor.org/terms (1958).

37. Breiman, L. *Random Forests*. vol. 45 (2001).

38. Cortes, C., Vapnik, V. & Saitta, L. *Support-Vector Networks*. *Machine Leaming* vol. 20 (1995).

39. Pedregosa, F. *et al. Scikit-Learn: Machine Learning in Python*. *Journal of Machine Learning Research* vol. 12 http://scikit-learn.sourceforge.net. (2011).

40. Tan, M. & Le, Q. V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *International conference on machine learning* (2019).

41. Howard, A. G. *et al.* MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. (2017).

42. Russell, B. C., Torralba, A., Murphy, K. P. & Freeman, W. T. LabelMe: A Database and Web-Based Tool for Image Annotation. *Int. J. Comput. Vis.* **77**, 157–173 (2008).

43. Amigo, J. M., Babamoradi, H. & Elcoroaristizabal, S. Hyperspectral image analysis. A tutorial. *Anal. Chim. Acta* **896**, 34–51 (2015).

44. Benes, E., Matejka, G. & Fodor, M. Near-infrared spectroscopy for comprehensive analysis of dark chocolate composition. *Food Chem.* **469**, (2025).

45. Zhao, W. & Du, S. Spectral-Spatial Feature Extraction for Hyperspectral Image Classification: A Dimension Reduction and Deep Learning Approach. *IEEE Transactions on Geoscience and Remote Sensing* **54**, 4544–4554 (2016).