Published in Journal of Near Infrared Spectroscopy
30(3):107–121

Publisher version available at
DOI: 10.1177/09670335221078356

# Hyperspectral imaging for classification of bulk grain samples with deep convolutional neural networks.

**Erik Schou Dreier[1,2,3], Klavs Martin Sørensen[1], Toke Lund-Hansen [3], Birthe Møller Jespersen[1], and Kim Steenstrup Pedersen[2,4].**

**Abstract**

Near Infrared hyperspectral images (HSI) offers a fast and non-destructive method for seed quality assessment through combining spectroscopy and imaging. Lately, Convolutional Neural Networks (CNN) have shown to be promising tools for RGB image or spectral cereal classification. This paper studies how to design and implement deep CNN models capable of utilizing both the spatial and spectral dimension of HSI data simultaneously for analysis of bulk grain samples with densely packed kernels. Classification of 8 grain samples, including 6 different wheat varieties, were used as a test case.

The study shows that the CNN architecture ResNet, originally designed for RGB images, can be adapted to use the full spatio-spectral dimension of the HSI data through adding a linear down sample layer prior to the conventional ResNet architecture. Using traditional spectral pre-processing methods before passing the data to the CNN does not improve the classification accuracy of the networks, while a channel-wise image standardization improves the accuracy significantly. The modified ResNet applied to the full spatio-spectral dimension has a classification accuracy of up to $99.75 \pm 0.02$ %, outperforming both purely spectral ($86.5 \pm 0.1$ %) and purely spatial ($98.70 \pm 0.01$ %) based methods in terms of accuracy, indicating that utilizing spatio-spectral correlation can improve sample classification, but also that grain classification is primarily solved using spatial information.

The findings reported in this paper demonstrate how CNN networks can be designed to leverage spatio-spectral information in hyperspectral data. The combination of HSI and spatio-spectral CNN networks shows a possible method for fast prediction of bulk grain quality parameters where both spectral and spatial properties of the grains are important.

**Keywords**

Hyperspectral imaging, NIR spectroscopy, Artificial Intelligence, Convolutional Neural Networks, Bulk grain sample classification

## Introduction

Near Infrared (NIR) hyperspectral imaging (HSI) offers a fast and non-destructive method for seed quality assessment through combining spectroscopy and imaging [1, 2, 3, 4, 5, 6, 7]. However, HSI produces vast amounts of data, and it is difficult to extract useful information in situations where the combination of spectral and spatial properties could be of importance. In consequence, deep learning methods has gained increased interest, in particular within remote sensing applications, but lately also within other HSI application such as seed quality assessment [8]. In particular, Convolutional Neural Networks (CNN) are powerful tools for hyperspectral image classification, as they can be trained to extract important features in multidimensional datasets which through supervised learning methods can be used to solve classification tasks. Conventionally, deep learning methods have typically been applied to hyperspectral images (sometimes referred to as hypercubes) of single kernel or hyperspectral images of sparse distributed kernels so far from each other that single kernel segmentation can easily be done [9, 10, 11, 12, 13]. While this offers a good method for single kernel quality assessment, it requires more complex sample preparation where kernel separation is handled mechanically or manually. On the other hand, evaluation of bulk grain samples is more difficult, but the method has a potential for screening large amounts of grains rapidly, and to become a critical in-line quality assessment tool for mills, replacing the traditional grab-sampling QC methods. Bulk grains here refers to densely distributed kernels placed to close for robust segmentation of single kernels.

A frequently used example of the application of hyperspectral imaging within seed evaluation is the classification of grain varieties. Grain variety is an important parameter in grain pricing [14] and non-destructive variety determination has historically been done using either spectral or spatial properties of bulk cereal samples using traditional chemometric analysis of kernel spectra [4, 15], or to some

[1] Chemometrics and Analytical Technology, Department of Food Science, University of Copenhagen, Rolighedsvej 26, DK-1958 Frederiksberg, Denmark.
[2] The Image Section, Department of Computer Science, University of Copenhagen, Universitetsparken 1, DK-2100 Copenhagen, Denmark.
[3] Foss Analytical, Foss Allé 1, DK-3400 Hillerød, Denmark.
[4] Digital Collections, Natural History Museum of Denmark, University of Copenhagen, Universitetsparken 15, DK-2100 Copenhagen, Denmark.

**Corresponding author:**
Erik Schou Dreier
Email: erik.dreier@food.ku.dk

extent using machine-vision on gray scale or RGB color images [16]. The possibility of classification using either spectral or spatial properties clearly illustrates grain variety determination as a spatio-spectral problem. This has let to attempts of using deep learning network with feature fusion of both spectral and spatial properties of rice [11], or using linear and quadratic discriminators together with wavelet transforms to determine wheat varieties from hyperspectral images [17]. The use of CNNs for HSI classification has to our knowledge so-far been limited to images of single kernels and primarily only used for spectral feature extraction with 1D CNN [10, 13] and shallow 2D networks for spatial feature extraction across various image channel [12, 18]. However, deeper neural network such as VGG16 has been applied to similar data-structures as HSI, such as sensor fusion of short wave infrared, visible (RGB), and visible-near infrared, where it has shown promising results [19].

This paper examines, to the best of our knowledge as the first, the possibility of using deep CNNs directly on the full 3 dimensional HSI data for analysing bulk grains. Here we use grain variety classification as test case for this form of analysis. The scope of this research is to identify approaches for and investigate potential complication of applying deep CNNs, conventionally used for RGB images, to the multi-channeled hyperspectral images using both spatial and spectral information simultaneously. The paper does not attempt to show improved performance compared to other methods of grain classification. Two adaptations of the network design ResNet[20] which can combine both spectral and spatial properties of hyperspectral images are investigated for grain classification: The first, is a modified version of a conventional 2D-ResNet with 2 dimensional convolutional kernels where an additional linear down-sampling layer is added to make use of the spectral dimension. 2D-ResNet has been a benchmark network within RGB-image classification since first proposed. It uses residual learners which have proven effective in reducing problems with vanishing gradient in deep neural network training, which otherwise made it difficult to train deep neural networks [20]. The second network design is a 3D-ResNet model with 3 dimensional convolutional kernels. The use of 3 dimensional convolutional kernels have previously been proposed for HSI within remote sensing [21] as a possible method to utilize spatio-spectral information simultaneously.

The study investigates the importance of spectral pre-processing and methods for effective spectral down-sampling prior to passing data to deep CNN networks. The robustness of the methods is evaluated through using varying moisture content across training, validation, and test sets, and distorting the images though small adjustment of the image setup. Furthermore, the results of the ResNet models HSI classification are compared to the performance of purely spectral and spatial classification to highlight the spatio-spectral properties of the grain variety classification. The results based on spectral classification is performed with Partial Least Squares Discrimination Analysis (PLS-DA) and Support Vector Machine (SVM), and purely spatial classification is made by applying 2D ResNet to gray scale NIR images.

## Sample preparation

In order to study the use of CNN on hyperspectral images of grain, 8 different grain samples were chosen of between 134 and 192.5 g each. Of these, 4 were conventional wheat varieties from 2018 of which 1 was a spring wheat and 3 were conventional winter wheat varieties of high baking quality (Type A) and low/medium baking quality (Type B). 3 wheat varieties were harvested in 2019, of which 2 were organic spring wheat heritage varieties, named Halland and Øland wheat and one an organic spelt (dinkel wheat). The last grain sample was a Midsummer Rye, which was included to impose a larger variance in kernel shape and size. The 8 samples were all stored under similar dry conditions for at least a year.

All 8 samples were divided into training, validation, and test sets for the purpose of experimentation with CNN models. For all grain varieties, about 27 g were selected for the validation set and 39 g for the test sets. Information of the three data sets can be seen in table 1. After dividing the grain samples into training, validation, and test sets, the three sets were stored independently in zip-locked bags for 4 months.

All samples were measured in a random order, where for each training sample, 10 images of densely packed kernels were acquired followed by 6 images of sparsely packed kernels. For both the validation and test sets, 5 dense images followed by 3 sparse images were acquired. The amount of grains within one dense image varied between 8 g and 12 g, amounting to between 150 and 600 kernels per image, depending on grain variety. The sparse images contain about 50 to 100 kernels depending on the size of the grain kernel type. The amount of kernels per training sample, only allowed for 4-7 unique dense images per grain variety sample. To obtain 10 dense images of each variety, the kernels were remixed after all kernels had been images ones, and additional images were taken of the remixed kernels until reaching 10 dense images. The same procedure was applied to the validation and test set to obtain 5 dense images as only 2-4 unique dense image could be obtained. This image sequence was repeated 3 times (labeled $S_1$, $S_2$, and $S_3$), each time having a different random order of samples, on three separate days. The test set was only measured on the first and third measurement day. This procedure was chosen to minimize the effect of image variation due to heat or other non-sample related variations caused by the imaging setup. Between the second and third measurement day, a small adjustment of the hyperspectral camera position and focus was made to introduce variations in the the experimental setup. The chosen method results in every single kernels being images at least three times (two for the test set), however, as images contain multiple kernels, all hyperspectral images are expected to contain a unique composition of kernels. An illustration of the experimental protocol is shown in figure 1(a).

The effect of kernel moisture on the grain classification accuracy has previously shown to be important when using machine vision on RGB images of kernels [22]. To control the humidity of the grains during the experimental sequences and evaluate differences in protein content of the grains, moisture and protein content were measured using the grain analyser Infratec NOVA after the second ($S_2$) and third

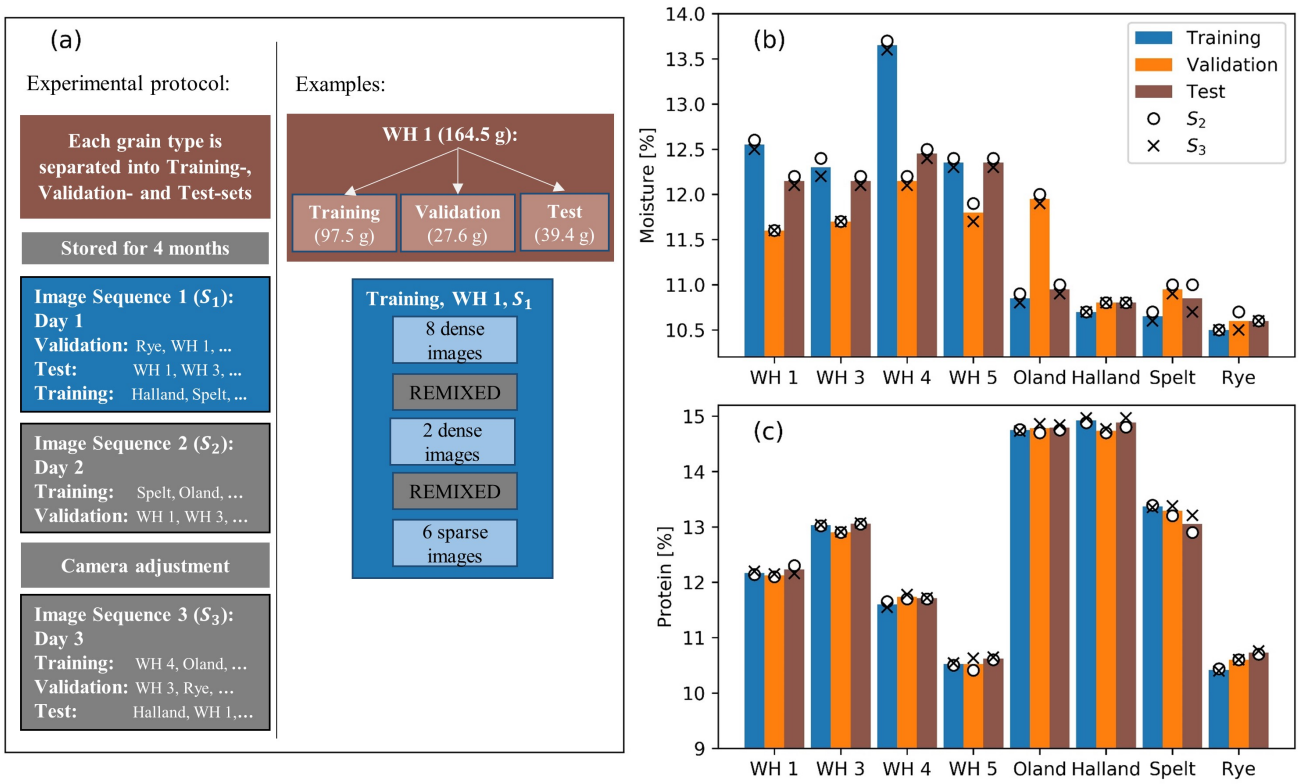


**Figure 1.** Experimental protocol and dataset moisture and protein composition. (a) Illustration of the experimental protocol as explained in the main text with examples of the data-set separation and image acquisition for the WH 1 sample. Average moisture (b) and protein (c) content of the 8 grain samples for the three separate datasets: training, validation and test. The measured moisture and protein values after the second ($S_2$) and third measurement sequence ($S_3$) are shown with black markers. The grain sample properties are shown in table 1.

($S_3$) measurement sequence – See moisture and protein percentages in figure 1(b,c). The variation in humidity in-between data-sets and grain types is a result of storing grain samples in separate bags for 4 months.

## Hyperspectral Image acquision

The imaging setup is shown in figure 2. An FX17 line scan camera with an InGaAs sensor from Specim was placed 130 mm above a conveyor belt running at a speed of about 0.05 m/s. The FX17 camera has 640 spatial pixels and is sensitive in the range 900-1700 nm with 224 evenly distributed spectral channels. For the results presented in this paper, the data was acquired at a frame rate of 350 frames per seconds. To standardize the image acquisition, the grain kernels were placed in a sample tray. In-front and in-back of the sample tray, 1 cm wide strips of white PTFE (Polytetraflourethylen) foil were placed for white reference measurements. Two checkerboards were placed in-front and in-back of the tray and used to measure and standardize the image pixel size along and perpendicular to the propagation direction of the sample tray. The pixels size along the grain propagation direction was measured to vary between 0.14 and 0.15 mm due to speed variations of the conveyor belt. The spatial resolution of the image system is $0.21 \pm 0.01$ by $0.23 \pm 0.02$ mm$^2$ parallel and perpendicular to propagation direction, respectively, as estimated from the smearing of the checkerboards. The resolution was slightly worse on the sides of the images. To illuminate the sample 6 halogen reflector spotlights were used, 3 placed in-front and 3 behind the camera seen along the direction of the conveyor. Two mirrors reflect the light from the halogen light bulbs onto the sample at an angle of $14.5 \pm 0.5$ degrees with respect to vertical, as measured from the length of the shadow of known height reference measurements.

All images were cropped after acquisition to contain only the sample tray with grains and the pixel size standardized using the checkerboards as reference. The acquired hyperspectral images were interpolated to ensure a uniform pixel size parallel and perpendicular to the propagation direction of $(0.15 \times 0.15)$ mm$^2$.

To correct for variations in light intensity, pixel inhomogeneity, camera dark current, etc, the raw HSI intensity spectra $I_{i,t}$ was transformed into absorbance spectra

$$A_{i,t} = -\log\left(R_{i,t}\right) = -\log\left(\frac{I_{i,t} - D_i}{W_i - D_i}\right), \qquad (1)$$

for each $i$'th spatial camera pixel at time slice $t$. Here $W_i$ and $D_i$ are the white and dark reference spectra for each spatial detector pixel, respectively, and $R_{i,t}$ the corresponding reflectance spectrum at time slice $t$. $W_i$ is calculated as the average recorded light intensity reflected off the white PTFE foil in front and back of the sample tray. The dark reference was measured as the dark current in the detector with closed shutter at the end of each measurement sequence. The first 10 and last 10 spectral channels of the hyperspectral image data was cropped out for all images, due to low detector sensitivity within this range, leaving the data with 204 spectral channels for all further analysis. All spectral

**Table 1.** Grain sample properties with grain type, country of cultivation, harvest year, 1000 kernel weight, and weight of training, validation and test-set for each sample. DK: Denmark, SE: Sweden

| Label | Grain type | Country | Year | 1000 kernel weight (g) | Training-set weight (g) | Validation-set weight (g) | Test-set weight (g) |
|---|---|---|---|---|---|---|---|
| Rye | Midsummer Rye | DK | 2019 | 19.5 | 75.0 | 31.1 | 39.3 |
| Spelt | Spelt wheat | DK | 2019 | 46.5 | 62.5 | 26.3 | 39.1 |
| Halland | Halland wheat | DK | 2019 | 26.5 | 61.6 | 26.2 | 39.3 |
| Oland | Øland wheat | DK | 2019 | 29.3 | 126.5 | 26.6 | 39.4 |
| WH 1 | Winter wheat, Type A | SE | 2018 | 43.1 | 97.5 | 27.6 | 39.4 |
| WH 3 | Spring wheat | SE | 2018 | 42.8 | 81.5 | 27.3 | 39.4 |
| WH 4 | Winter wheat, Type A | DK | 2018 | 45.2 | 112.3 | 27.4 | 39.4 |
| WH 5 | Winter wheat, Type B | SE | 2018 | 41.2 | 67.3 | 27.3 | 39.4 |

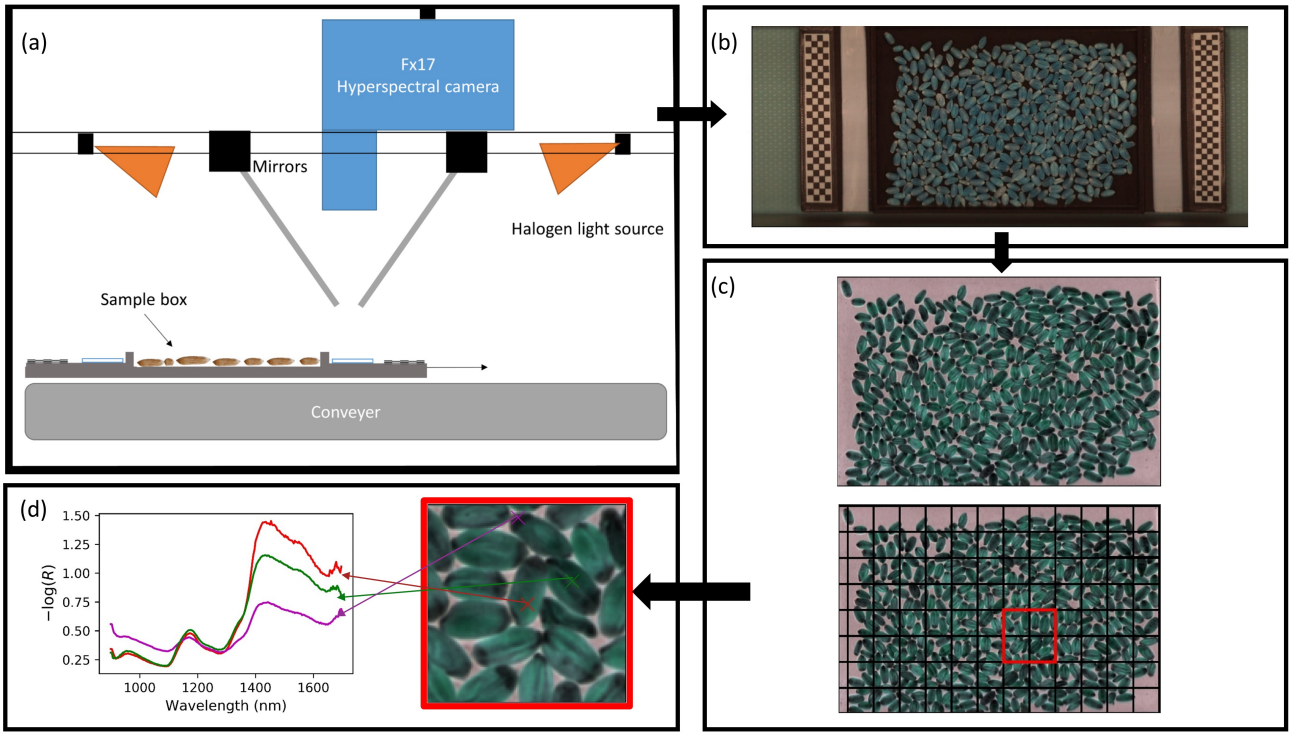


**Figure 2.** Imaging setup used to acquire hyperspectral images. (a) The setup consists of a linescan, hyperspectral FX17 camera placed above a moving conveyor belt. Two sets of each 3 halogen reflector spotlights were placed in front and back of the camera seen in the propagation direction of the conveyor. Two aluminum mirrors were used to decrease the incoming angle of light seen from vertical to minimize the effect of shadows. (b) The grain samples are presented to the camera in a tray. In front and back the tray a strip of PTFE foil and a checkerboard pattern were placed for white reference and pixel size standardization. (c) The part of the image containing grain samples were segmented and separated into image windows of 128 by 128 pixels with an overlap of 64 pixels. (d) Each window is a hyperspectral image with 128x128 spatial pixels and 224 wavelength channels yielding a NIR infrared spectrum in each pixel.

absorbance images have been made public available in online repositories with the full 224 channels, see Data availability statement.

A sliding window approach was used to train and validate the classification models, as illustrated in figure 2. The sliding window crops the hyperspectral kernel images into smaller windows of 128 by 128 pixels with an overlap of 64 pixels. The sliding window approach has three purposes: It allows for a spatially resolved grain classification on the size of up to about 10-15 kernels per window, thereby utilizing the potential for spatially resolved classification. The smaller images are better suited for CNN networks, as less memory is needed while passing the hypercube through the network. Lastly, the small image windows allow for more images to train and evaluate the classification results. To minimize the data size further and allow for faster training of the models, the number of spectral image channels was reduced to a maximum of 68 channel through averaging every third spectral channel. This correspond to a spectral wavelength bin size of 10.7 nm, around 30 % larger than the actual expected spectral resolution of the FX17 of 8 nm (FWHM) according to manufacture specifications. Below, the use of additional spectral down-sampling through further binning or using Principal Component Analysis (PCA) is investigated.

We define the kernel area ratio between pixels containing kernels and the total number of pixels in the sliding windows as a model hyper-parameter

$$\rho_P = \frac{\text{\# of pixels with kernels}}{128 \times 128}. \quad (2)$$

For all results presented later, unless stated otherwise, a threshold on the kernel area ratio of $\rho_P = 0.5$ was used, i.e. requiring that 50 % of the image window contains kernels. The pixels containing kernels were segmented using an Otsu threshold on the mean hyperspectral images.

The window size of 128 by 128 pixels and $\rho_P = 0.5$ resulted in 16666 windows for training, 6536 for validation and 4322 for testing. As the grain kernel images should be invariant under rotation and flip, rotation and flipping of the spatial dimension is used as data argumentation on the training data. Thereby, the number of training image windows were increased by a factor of 8 to 133328 in total.

## Models

Two deep convolutional neural networks were evaluated: One slightly modified version of a 2D ResNet network and a 3D ResNet model as illustrated in figure 3. ResNet uses residual learners which computes residual features by introducing a skip connection for each basic block, as illustrated in figure 3(c), where the input image is added to the image after passing through the convolutional layers. It should be noted that the standard ResNet uses a spatial image size of $256 \times 256$ pixels, whereas a spatial image size of $128 \times 128$ is used in this paper. Similarly to the original ResNet proposed by He et al. [20], the ResNet models used here are implemented with Batch Normalization (BN)[23] following each convolutional layer and Rectified Linear Unit activation function (ReLu)[24]. Due to the relatively low number of training images, this study uses an 18 layer deep ResNet model, although better accuracy may possibly be obtained with deeper models.

We denote the modified 2D ResNet18 network used in this paper as 2D ResNet18-spec. In order to make use of the spectral dimension efficiently, and allow the network to do a quick down-sampling of the data, we introduce a modification of the conventional ResNet18, by adding a 1x1x(number of spectral channels) convolution layer with 3 output channel (i.e. the layer contains 3 filters) in front of the first ResNet layer, thereby increasing the number of layers to 19. This convolution layer enables the network to perform a linear downs-sampling of the hyperspectral image data, by a linear combination of spectral bins and reducing this the three channel input required by the conventional ResNet.

The 3D ResNet18 uses 3D convolution kernels, which means that spectral dimension of the dataset is seen as a 3rd dimension in the data instead of image channels. 3D convolutional layers allows for a volumetric comparison of signal changes across the spectral-spatial data-cube, which could be important for classification. This is unlike 2D filters, which treats each channel independently. The network is based on work by Hara et al.[25], with a slight modification which means that the network only takes 1 input channel instead of the 3 used for RGB-classification. For the 3D convolutions, stride and max-pooling is done across all three dimensions, both spectral and spatial. The feature channels created as the hypercube is passed through the network here represents a 4th data dimensions, meaning that just before the final average pool, the 3D ResNet has reduced a 128x128x68 data cube to 512 cubes of 4x4x2, each representing a spatio-spectral feature.

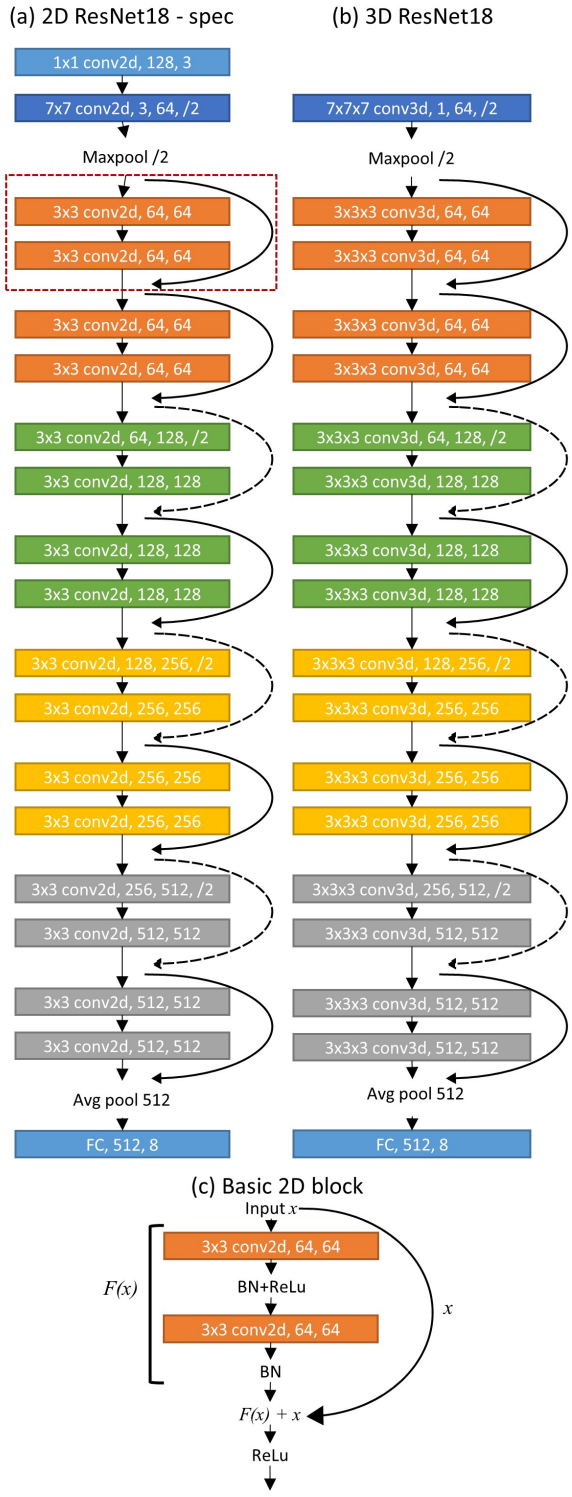


**Figure 3.** Illustration of the 2D ResNet18-spec (a) and 3D ResNet18 (b) models used for grain classification. Each colored box shows a convolutional layer, with the first numbers giving the spatial kernel size (spatio-spectral for 3D ResNet), and the last two indicating the number of input and output channels, respectively. '/2' shows where a stride of 2 is used to down-sample the spatial dimension (spatio-spectral for the 3D network). The semi circle lines show skip connection, with dashed lines showing further down-sampling of the input. The dashed red rectangle highlights a basic block in the 2D ResNet model which is shown in more details in panel c. BN is Batch normalization and ReLu is a Rectified Linear Unit activation function.

### *Model implementation*

Similar to the original implementation of ResNet [20], we use stochastic gradient decent as optimizer, Kaiming weight initialization [26], a weight decay of 0.0001 and a momentum of 0.9. We use cross entropy loss function and initialize with a learning rate of $2 \cdot 10^{-2}$ which is divided by 10 when the error plateaus with a patience of 6 epochs. The initial learning rate was found by scanning the loss as function of the learning rate and finding the learning rate where loss diverges. The method is similar to the one proposed in reference [27]. A scan of model accuracy as function of batch size size showed significant improvement of the models' accuracy with decreasing batch size. A difference in performance of more than 2 % was observed between a batch size of 128 and 16. Larger batch size, however, significantly improves computation time. The uncertainty in accuracy of the models were evaluated through training three independent, but identical, models on three different data partitions each containing 80 % randomly selected image windows from the available training-set. The uncertainty of the model accuracy is in the following defined as the standard deviation of the three models results.

## Evaluation of data pre-processing and spectral down-sampling

To efficiently use deep learning models on the vast hyperspectral dataset, down-sampling of the spectral dimension can be useful prior to passing the data to a deep neural network. Two methods for down-sampling the spectral dimension before passing to the ResNet 2D model are investigated in the following - PCA and simple binning of the spectral channels. For the PCA approach, the score images corresponding to the 10 first principal components were selected. The score images were computed through independently applying the PCA across all spectra in each hyperspectral grain image before cropping $128 \times 128$ windows. For binning, a varying number of spectral bins was investigated from 1 to 68 bins across the wavelength range of the FX17 camera. For values smaller than 68 bins, data is re-interpolated from the 68 bins data for computational efficiency. The 3D ResNet model was only tested on data which is binned to 68 bins, as the network architecture is designed to work across the spectral dimension.

Image standardization may be useful in mitigating channels giving extra weight from large data values and the risk of exploding/vanishing gradient. In the following, we also investigate different approaches for spectral data standardization prior to passing the data to the network.

In figure 4(a), we compare the performance of 2D and 3D ResNet models om raw hyperspectral image data with different standardization methods. For the 2D ResNet model, both down-sampling through reducing the spectral dimension to the first 10 principal components as well as down-sampling through binning to 40 bins were tested. For 3D ResNet model the data was fixed at 68 bins which is the maximum number of bins passed to the network. For this scan, 3 models were independently trained for each data-point on 40 % of the image data with a batch size of 128 for 10 Epochs.

The standardization methods in figure 4(a) is done on an image window basis in two steps described by the number pair $[N_1, N_2]$. First, the pseudo absorption images data $A(x, y, c)$ is standardized using methods described by $N_1$ so $A(x, y, c) \xrightarrow{f(N_1)} A^*(x, y, c)$ followed by a second step $A^*(x, y, c) \xrightarrow{f(N_2)} A^\dagger(x, y, c)$. $A^\dagger(x, y, c)$ is then passed to the network. Down-sampling of the spectral dimension to less than 68 channels is done in-between the two steps. If $f(N_1 = 0)$ no standardization is done in the first step. For $f(N_1 = 1)$, each channel is mean centered and auto-scaled by subtracting the mean of each spectral channel

$$\bar{A}_{x,y}(c) = \frac{1}{N_x N_y} \sum_{x,y}^{N_x, N_y} A(x, y, c), \tag{3}$$

and normalizing with the standard deviation of each spectral channel

$$\sigma_{x,y}(c) = \sqrt{\frac{1}{N_x N_y} \sum_{x,y}^{N_x, N_y} \left(A(x, y, c) - \bar{A}_{x,y}(c)\right)^2}. \tag{4}$$

Here, $N_x$ and $N_y$ are the number of pixels in the two spatial dimensions. If $f(N_1 = 2)$ each pixel spectra is mean centered and auto-scaled by subtracting the mean of each pixel's spectrum

$$\bar{A}_c(x, y) = \frac{1}{N_c} \sum_{c}^{N_c} A(x, y, c), \tag{5}$$

and normalize with standard deviation of each pixel's spectrum

$$\sigma_c(x, y) = \sqrt{\frac{1}{N_c} \sum_{c}^{N_c} \left(A(x, y, c) - \bar{A}_c(x, y)\right)^2}, \tag{6}$$

corresponds to a Standard Normal Variate (SNV) correction. The three different methods can be summarized accordingly

$$\begin{aligned} N_1 = 0: &\quad A^*(x, y, c) = A(x, y, c) \\ N_1 = 1: &\quad A^*(x, y, c) = \frac{A(x, y, c) - \bar{A}_{x,y}(c)}{\sigma_{x,y}(c)} \\ N_1 = 2: &\quad A^*(x, y, c) = \frac{A(x, y, c) - \bar{A}_c(x, y)}{\sigma_c(x, y)} \end{aligned} \tag{7}$$

The $N_2$ number represents the next step in hyperspectral image standardization where $f(N_2 = 1)$ corresponds to re-scaling the local patch with the minimum $\min_{x,y,c}(A^*)$ and maximum of the hyperspectral data $\max_{x,y,c}(A^*)$ across all data dimensions and $N_2 = 2$, equals re-scaling the local patch with the minimum $\min_{x,y}(A^*)(c)$ and maximum $\max_{x,y}(A^*)(c)$ of each channel accordingly

$$\begin{aligned} N_2 = 1: &\quad A^\dagger(x, y, c) = \frac{A^*(x, y, c) - \min_{x,y,c}(A^*)}{\max_{x,y,c}(A^*) - \min_{x,y,c}(A^*)} \\ N_2 = 2: &\quad A^\dagger(x, y, c) = \frac{A^*(x, y, c) - \min_{x,y}(A^*)(c)}{\max_{x,y}(A^*)(c) - \min_{x,y}(A^*)(c)} \end{aligned} \tag{8}$$

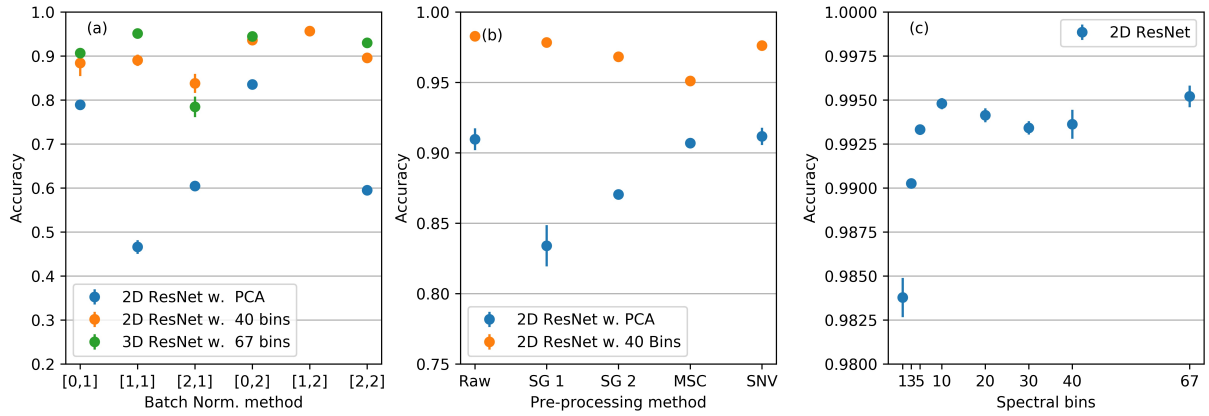


**Figure 4.** Average best accuracy found by three independently trained models as function of data pre-processing, image standardization, and spectral down-sampling prior to providing the data as input to the two different network types. (a) The best accuracy of 2D and 3D ResNet models after 10 epochs as function of image standardization method of raw data, represented with numbers $[N_1, N_2]$ in the axis tick labels. $N_1$ and $N_2$ correspond to different methods for standardization of the hyperspectral images as described in equation 7 and 8. (b) 2D ResNet model accuracy as a function of spectral pre-processing methods using spectral binning and PCA to down sample the data. (c) Performance of 2D ResNet model on raw data as function of the number of spectral bins. The spectral down-sampling from 68 to the specified number of bins is done between the two standardization steps through average.

It should be noted that for binning of 68 and PCA results the two standardization follows directly after each other with no in-between data-processing, hence combination [2,1] is excluded as it is similar to [0,1] without further data down-sampling between the two steps.

Figure 4(a) shows that normalization [1,2] gives an accuracy of more than 95 % which is significantly higher than any of the other standardization methods used for the 2D ResNet network when applied to 40 bins data. For PCA down-sampled data, the results are slightly worse than the binned approach, with an accuracy below 85 % across all standardization methods. The best result is obtained with [0,2]. For 3D ResNet, [1,1] standardization seems to perform slightly better than other approaches. However, a channel-wise normalization, such as [0,2], also works well for 3D ResNet, which means that a channel-wise normalization similar to what is done with RGB images perform very well across different networks and down-sampling methods.

Conventionally, chemometrics modeling of NIR spectra uses different correction algorithms prior to analysing the data to mitigate of light scattering artifacts [28], primarily from Rayleight scattering occurring when the radiation interacts with particles whose dimensions are small compared with the wavelength of the light. Here, we investigate 4 different correction methods which are traditionally used for NIR spectra: 1st and 2nd derivative using a Savitzky–Golay (SG)[29], multiplicative scattering correction (MSC)[30] and standard normal variate (SNV)[31]. For both SG filters, a second order polynomial fitting was chosen with a kernel size of 13 for the 1st order derivative and 17 for the second order derivative. The filter sizes was determined from the best performance of the SVM algorithm described in the section Importance of spectral and spatial dimension, below. For MSC, the reference spectrum required to correct the spectrum was found by computing the average spectrum of all grain kernels in a sample type. The average grain kernel spectrum was found through segmenting kernels from background in the images and averaging. All methods were applied to full rank spectral data with 224 spectral channels before down-sampling to the first 10 PCA components explaining the largest variance or binning to 68 channels. The summed explained variance of the first 10 PCA components were for SNV 98.0 %, for MSC 99.6 %, for SG 1st derivative 97.9 %, and for SG 2nd derivative 89.9 %. Spectral corrections algorithms are only applied to data past into 2D ResNet models. It is assumed that channel-wise derivatives-like filters should be learnable for a 3D ResNet model, if relevant for the grain classification task, and hence only raw data was evaluated for this method. The best performing standardization method, as shown in figure 4(a) for the two down-sampling approaches, was applied prior to passing the hyperspectral image windows to the network.

The four spectral correction approaches are compared to the results without any correction in figure 4(b). The accuracy are based on 3 models trained on each 80 % of the training data with 128 mini-batch size. The best performance is for both binning and PCA results found when not applying any corrections. Furthermore, it is noted again that PCA used for spectral down-sample performs significantly worse than simple binning.

Lastly the number of bins were evaluated on raw data for the 2D ResNet network to find the optimal number of spectral bins. For these results, 3 models were trained for 15 Epochs on 80 % of the training data for each bin-size with a mini-batch size of 16. For all bin-sizes, the image standardization [1,2] was used. The results are shown in figure4(c) and show a tendency to yield slightly better results with 68 bins, but a smaller number of spectral bins, such as 10, almost performs equally well.

## Importance of spectral and spatial dimension

The performance of the ResNet models are compared with the results obtained from purely spatial and spectral data in

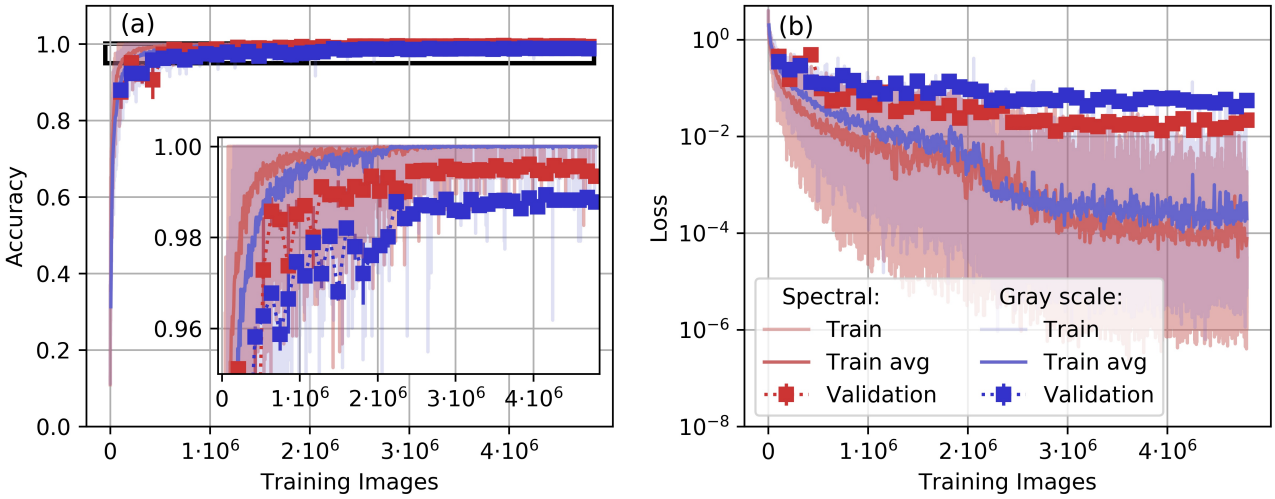


**Figure 5.** Average accuracy (a) and categorical cross entropy loss (b) across three independently trained models for 45 epochs on 68 channel spectral image and gray scale image data. The insert in panel (a) is a zoom in on 0.95-1 accuracy. The validation markers show the mean accuracy for each epoch of the three independently trained models.

**Table 2.** Highest average accuracy of 3 independently trained 2D ResNet18-spec models trained on 68 channel spectral images and gray scale images, and the 3 independently trained 3D-ResNet18 models trained on 68 channel spectral images. The 2D ResNet18 models were trained for 45 epochs, and the 3D-ResNet18 for 15 epochs. For comparison the best accuracy after 15 epochs is reported for all models. The results are compared to the accuracy of SVM and PLS models trained on average spectral data. The highest accuracy for the different data-sets is highlighted in bold. Batch standardization method described by the number pair $[N_1, N_2]$ is given by equation 7 and 8. SG: Savitzky-Golay.

| Type | Pre-procesing | Batch standardization | Number of Epochs | Training [%] (N= 106662) | Validation [%] (N= 6536) | Test [%] (N=4322) |
|---|---|---|---|---|---|---|
| PLS | SG 1. derivative | - | - | $91.6 \pm 0.1$ | $73.78 \pm 0.04$ | $79.9 \pm 0.1$ |
| SVM | SG 1. derivative | - | - | $97.5 \pm 0.1$ | $85.6 \pm 0.2$ | $86.5 \pm 0.1$ |
| 2D ResNet Gray scale | Raw data | [0,1] | 15 (45) | $99.46 \pm 0.05$ ($\mathbf{99.99 \pm 0.01}$) | $98.4 \pm 0.1$ ($99.02 \pm 0.01$) | $98.1 \pm 0.2$ ($98.70 \pm 0.01$) |
| 2D ResNet 68 channels | Raw Data | [3,2] | 15 (45) | $99.61 \pm 0.04$ ($99.85 \pm 0.04$) | $99.5 \pm 0.1$ ($\mathbf{99.65 \pm 0.02}$) | $99.6 \pm 0.1$ ($\mathbf{99.75 \pm 0.02}$) |
| 3D ResNet 68 channels | Raw Data | [3,1] | 15 | $99.72 \pm 0,04$ | $99.3 \pm 0.1$ | $99.4 \pm 0.1$ |

table 2. The spatial-only results are obtained by applying the same 2D ResNet18-spec network used for hyperspectral images to a single gray scale NIR image of the kernels. The gray scale NIR images are generated by averaging the HSI data along the spectral direction without any pre-processing, simulating a "non-spectral resolved" NIR image acquired with an InGaAs detector. Thereby, it is possible to get a direct comparison of gray scale NIR images against spectral NIR images on the exact same data. The gray scale images are re-scaled to have values between 0 and 1 on a window to window basis with max-min scaling. For both spectral and gray-scale data, the 2D-ResNet models were trained on the full training set for 45 epochs using a mini-batch size of 16. Due to a 3 times longer training time, the 3D ResNet model was only trained for 15 epochs. Three models were independently trained for each network using randomly chosen 80 % of the training data for each model.

The hyperspectral image windows can also be averaged along the spatial directions leaving only a single spectrum. To avoid a background contribution to the average spectrum, the kernels are segmented through a Otsu threshold operation [32] before averaging. The resulting spectrum is equivalent to a NIR reflection spectrum of a few kernels and allows us to compare the performance using only spectroscopic information of the exact same kernels. The variety classification based on the spectral data-set was done using Partial Least Squares Discrimination Analysis (PLS-DA) and Support Vector Machines (SVM) with Radial Basis Function (RBF)-kernel. To optimize classification based on the linear PLS and the non-linear SVM models, the same pre-processing approach prior to binning as described in relation to figure 4 was attempted. For all results, a down-sampling to 68 bins was applied through averaging the spectral channels, similar to what was done for the spectral CNN results. The best result was found by applying Savitzky–Golay Filter 1st derivative with a kernel size of 13 before averaging. Due to the known difference in moisture between validation, test, and training sets, removing parts in the NIR spectrum was attempted.

The best performance of the SVM model was found by truncating the spectrum from 67 wavelength channels to

35, through removing signal below 970 nm, between 1100-1230 nm, between 1280-1470 nm, and above 1650 nm which coincides with the expected NIR absorption signal from the CH and OH vibrations that are related to the moisture and starch content. A 5.5 % increase in validation accuracy was found when training and validating on a truncated dataset compared to a full spectrum model. It is noted that truncating the spectrum such that only parts of the NIR-spectrum linked to OH is removed reveal 4.2 % lower performance than the best performance. For the PLS model, the highest validation accuracy was found through removing wavelength channels with the least impact on validation performance, yielding a 3.5 % accuracy improvement compared with a model trained on full spectral rank. The 21 channels selected for the PLS model has wavelength centers at $\lambda = [962.5, 973.2, 983.9, 994.6, 1166.1, 1176.8, 1187.5, 1348.22, 1358.9, 1369.6, 1433.9, 1444.6, 1455.4, 1466.1, 1476.7, 1487.5, 1498.2, 1508.9, 1616.1, 1626.8, 1637.5]$ nm, corresponding to 1st and 2nd overtone NH and CH information. 5 models were independently trained for both PLS and SVM models each using 80 % partitions of the training data.

Table 2 shows that 2D ResNet model applied to hyperspectral data achieves the highest validation and test accuracy after 15 epochs and that the 3D ResNet model performs almost equally well. The ResNet model applied to purely gray scale NIR images has a significant lower accuracy comparable to the performance of the ResNet models applied to the hyperspectral data after 15 epochs. The spectral 2D ResNet18 -spec model has almost converged after 15 epochs and only small improvements are found when training for 45 epochs. ResNet18 -spec trained on gray scale images does improve significantly from training 45 epochs, but is still performs wors than the ResNet models training on spectral data. Interestingly, it is also seen that even using PLS and SVM models on spectral data alone, a classification accuracy of above 80 % can be achieved, similar to what has previously been found using spectral data for grain classification [4, 10].

A comparison of training and validation loss and accuracy as function of the number of training images passed to the network is shown in figure 5 for the 2D ResNet18-spec model trained on spectral and gray scale images. The figure shows that both models converges after around 20 epochs, but similarly to the results presented in table 2, it can be seen that the 2D ResNet18-spec model requires fewer epochs to reach an accuracy close to the highest achieved accuracy when trained on spectral data compared to gray scale images.

The confusion matrices after 15 epochs of the test set for the 2D gray scale and hyperspectral data shown in figure 6 reveals that both models have problems identifying classes within the conventional wheat types, WH 1, WH 2, WH 3, and WH 4, and within the two organic wheats, Øland and Halland. However gray scale results are more prone to error within the conventional and organic classes. After fully converge, the same result was found, however, with much fewer mistakes by both models.

The weights of the convolution kernels in the first layer which down-samples the hyperspectral 68 bin data to three spectral channels is shown for the three independently trained models in figure 7. The three kernels of each model, $W_1$, $W_2$, and $W_3$, are shown in order of similar appearance.

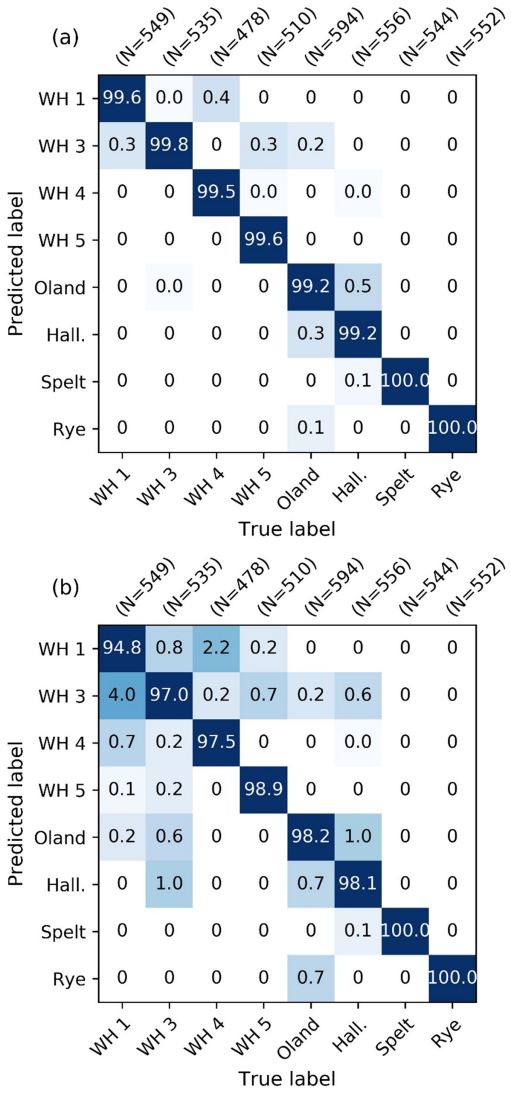


**Figure 6.** Confusion matrices for the test sets for 2D ResNet model applied to 68 bins spectral images (a) and gray scale images (b) showing the average best accuracy of 3 independent models after 15 epochs. The number of total image windows per class is given above the matrices.

As seen, the three models learn very similar features. Notice that moisture signature in the NIR signal is present around 975 nm and 1425 nm, which indicates that the model learns the importance of the moisture distribution in the kernels.

Using these weights will produce a 3 channel image as illustrated in figure 8.

## *Domain shift dependence*

Machine learning models, and in particular deep learning models, can be highly affected by domain shift between the trained data and the test data caused by changes to the imaging setup or sample presentation [33].

The three measurement sequences ($S_1$, $S_2$, and $S_3$) explained in Sample preparation section have in the previous presented results been combined for both training and validation of the models. In the following, the training and validation sets are evaluate on individual data sequences for the 2D ResNet18-spec model trained on both spectral images and gray-scale images as well as the SVM model trained

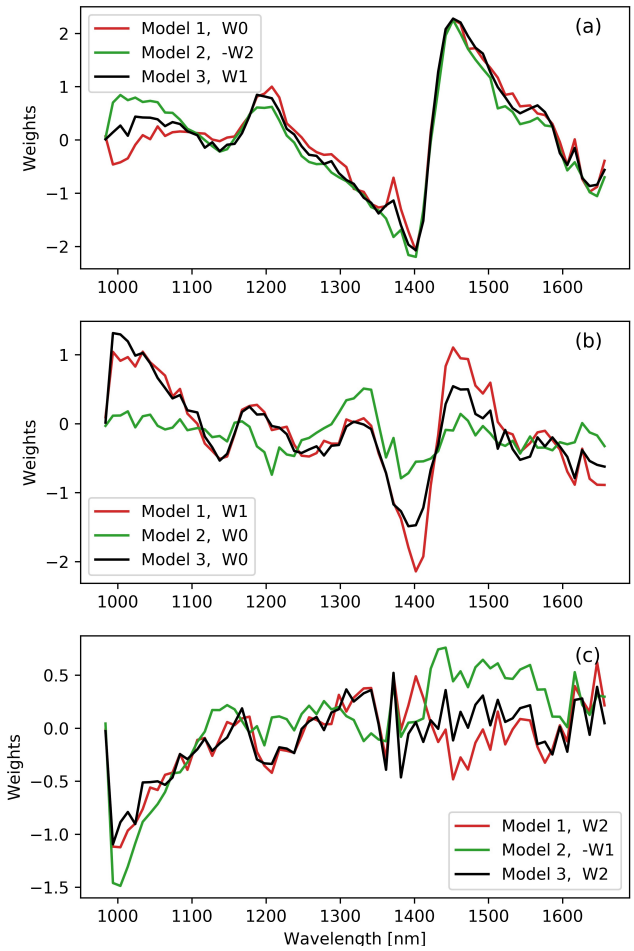


**Figure 7.** The weights of the 3 convolution kernels of the first layer (1x1 in ResNet18-spec, figure 3) of the three independently trained ResNet 2D models when trained on 68 bin spectral data. The order of the convolutional kernels has been reordered in accordance with qualitative similarity.

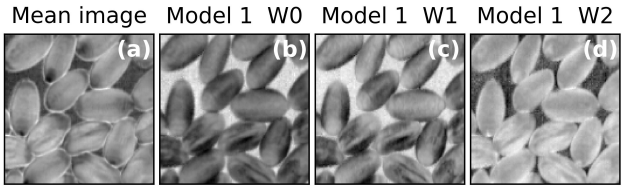


**Figure 8.** The result of applying the 1x1 2D convolution to 68 bin spectra data of the model 1. (a) mean gray scale image grain, (b) applying W0, (c) applying W1, and (d) applying W2.

on spectral data alone. The camera focus and position have intentionally been slightly adjusted between the $S_2$ and $S_3$ sequences to augment the data with physically introduced image variation. Table 3 shows the accuracy on the validation data from the $S_1$, $S_2$, and $S_3$ sequences, independently, when the three models are trained on training data from the first image sequence, $S_1$, the first and second, $S_1 + S_2$, and all three, $S_1 + S_2 + S_3$. For all values, three independent models were trained on randomly chosen 80 % of the accessible data and the remaining data is used as validation dataset. To ensure a similar number of training image batches for the ResNet models, the number of epochs were adjusted according to the number of data sequences included. The model for sequence $S_1$ was trained for 45 epochs, $S_1 + S_2$ for 23, and $S_1 + S_2 + S_3$ for 15 epochs.

A significant reduction in all models accuracy is observed in table 3 when validating the models on spectral images obtained with slightly modified setup than trained on. As expect, the biggest difference can be observed when training solely on the $S_1$ sequence and evaluating on $S_3$. Training on both $S_1$ and $S_2$ improves the prediction of $S_2$ and $S_3$ for all models. When including $S_3$ in the training data, the $S_3$ prediction accuracy is approaching the performance of the other two data-sets.

In general, the accuracy across all validation sets obtained with the spectral image ResNet model is the least affected by training on only one image sequence ($S_1$) compared to the accuracy when training on all ($S_1 + S_2 + S_3$). For the gray-scale image ResNet model and the spectral SVM model, a difference in $S_1$-validation accuracy of $2.9 \pm 0.2\%$ and $4 \pm 1\%$, respectively, is seen between models trained on $S_1 + S_2 + S_3$ and only $S_1$. For the ResNet models trained on spectral images, a reduction of only $0.7 \pm 0.2$ % is observed. The ResNet model trained on spectral images also shows the smallest difference on $S_3$-validation accuracy of just $3.2 \pm 0.1$ % between models trained on $S_1 + S_2 + S_3$ and only $S_1$. For the ResNet model trained on gray scale images the difference is $6.3 \pm 0.4$ %, and for the spectral SVM model a difference of $19 \pm 1$ % is observed.

NIR spectra are significantly affected by scattering [28], and the effect of slightly adjusting camera position and focus will not only affect the spatial dimensions but also change the spectral properties. As seen from the results obtained with the SVM model trained on spectral data in table 3, the spectral dimension is significantly affected by the change in camera position and camera focus. The reduction in accuracy between the validation sets is likely not caused by a change in grain humidity, as only a small variation in moisture is observed in the grain after the second and third image sequence in figure 1 compared to the differences between validation and training dataset. Furthermore, parts of the spectrum linked with OH (moisture) vibrations have been cropped from the spectrum before evaluating the SVM model. Notice that no spectral cropping was done for the ResNet models.

The neural networks are not only affected by changes to the experimental setup but also variation in how the grains are distributed in the images. Figure 9 illustrates this dependence by showing prediction maps of full images of densely and sparsely distributed kernels for both gray scale and spectral images. The prediction maps are created through evaluating the image with a sliding window (128x128 pixel) with step size of 16 pixels and finding the most probable prediction of all the models' prediction across overlapping windows within a pixel. The result is multiplied with a binary kernel segmentation mask to show only the kernels. The Resnet models are the same as the ones used for the results presented in table 2.

Figure 9 shows that the most probable prediction is generally accurate for the dense kernel images for both spectral and gray scale data. However, for the sparsely packed kernel image, both models makes multiple mistakes. This shows a problem with a domain shift between the image windows the models are trained on, which has a kernel area

**Table 3.** Accuracy as function of data acquisition sequences for the 2D ResNet18-spec model trained on spectral images for 15 epochs, 2D ResNet18-spec model trained on gray scale images for 15 epochs, and SVM models trained on spectral data. The models are trained on three different data-sets corresponding to data from the first acquisition sequence ($S_1$), first and second ($S_1 + S_2$), and all three sequences($S_1 + S_2 + S_3$). All models are validated on the three acquisition sequences $S_1$, $S_2$, and $S_3$, independently. Each accuracy value reported corresponds to the average best result of three independent models. The accuracy of all models trained on data from image sequence $S_1$ and validated on image sequence $S_3$ are underlined and the accuracy of models trained on $S_1 + S_2 + S_3$ and validated on $S_1$ is in bold to highlighting the difference.

| | | | Validated on: | | |
|---|---|---|---|---|---|
| 2D ResNet18 68 channels | Trained on: | | $S_1$ | $S_2$ | $S_3$ |
| | | $S_1$ | $99.0 \pm 0.1$ % | $98.7 \pm 0.1$ % | *95.9 ± 0.1 %* |
| | | $S_1 + S_2$ | $99.65 \pm 0.03$ % | $99.65 \pm 0.05$ % | $98.0 \pm 0.1$ % |
| | | $S_1 + S_2 + S_3$ | **99.7± 0.1** % | $99.80 \pm 0.05$ % | $99.06 \pm 0.03$ % |
| | | | Validated on: | | |
| 2D ResNet18 Gray scale | Trained on: | | $S_1$ | $S_2$ | $S_3$ |
| | | $S_1$ | $96.1 \pm 0.2$ % | $95.2 \pm 0.2$ % | *91.2 ± 0.4 %* |
| | | $S_1 + S_2$ | $98.8 \pm 0.1$ % | $97.8 \pm 0.1$% | $96.1 \pm 0.1$ % |
| | | $S_1 + S_2 + S_3$ | **99.0± 0.1** % | $98.6 \pm 0.1$ % | $97.5 \pm 0.1$ % |
| | | | Validated on: | | |
| SVM | Trained on: | | $S_1$ | $S_2$ | $S_3$ |
| | | $S_1$ | $81 \pm 1$ % | $84 \pm 1$ % | *65± 1 %* |
| | | $S_1 + S_2$ | $85.3 \pm 0.2$ % | $84 \pm 1$ % | $76.2 \pm 0.4$ % |
| | | $S_1 + S_2 + S_3$ | **85.7 ± 0.2** % | $87.1 \pm 0.4$ % | $83.6 \pm 0.2$ % |

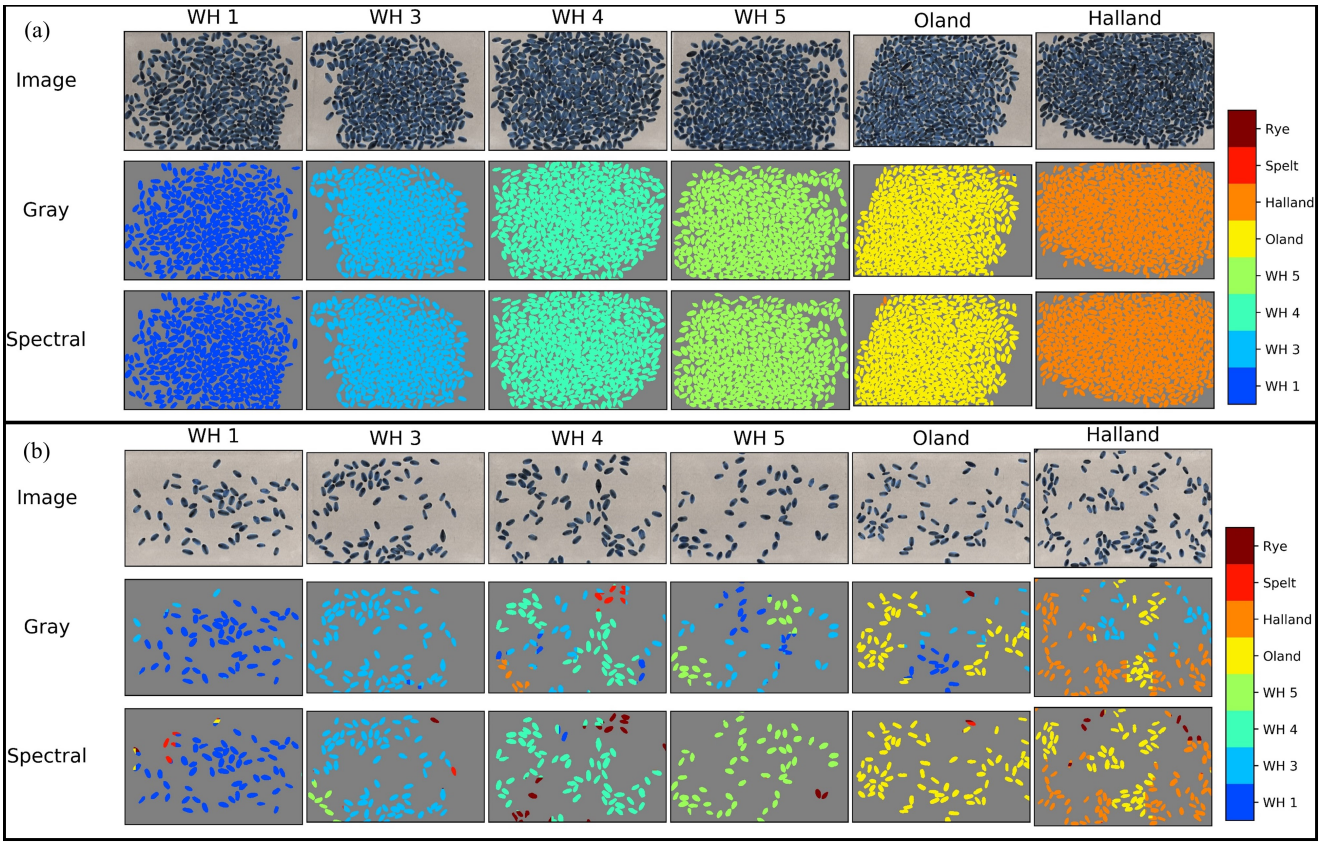


**Figure 9.** Prediction maps of a dense (a) and sparse (b) packed grain kernel datasets showing the most probable grain type as predicted by the 2D ResNet models trained and evaluated on either gray scale (Gray) or hyperspectral 68 channel (Spectral) data.

ratio (eq. 2) of more than 0.5, and the image windows from the sparse kernel image data where a significant number of windows have a kernel area ratio of less than 0.5. It should be noted that figure 9 shows single kernels which are classified as 2 or 3 different classes. This is an artefact of the window based classification. The CNN's could be trained to predict single kernel classes on the sparsely packed kernel images by employing a simple segmentation algorithm. Single kernel classification is, however, outside the scope of this paper which aims to show a method for generic classification of bulk grain samples where accurate single kernel segmentation is difficult or not needed. Hence the figure is merely intended to show problematic effect of domain shift between the densely packed kernels in the training images and the more sparsely packed kernels in the validation images.

In figure 10, this effect is quantify, by calculating the prediction accuracy of the models across all validation data as function of the kernel area ratio in each window, through separating the data into bins of $\rho_P = 0.1$, eq. 2. The figure clearly shows a drop in accuracy when decreasing the kernel area ratio of the validation images below the domain of the training images ($\rho_P > 0.5$). The result shows a clear domain dependence and illustrates the importance of sample preparation when training deep neural networks.

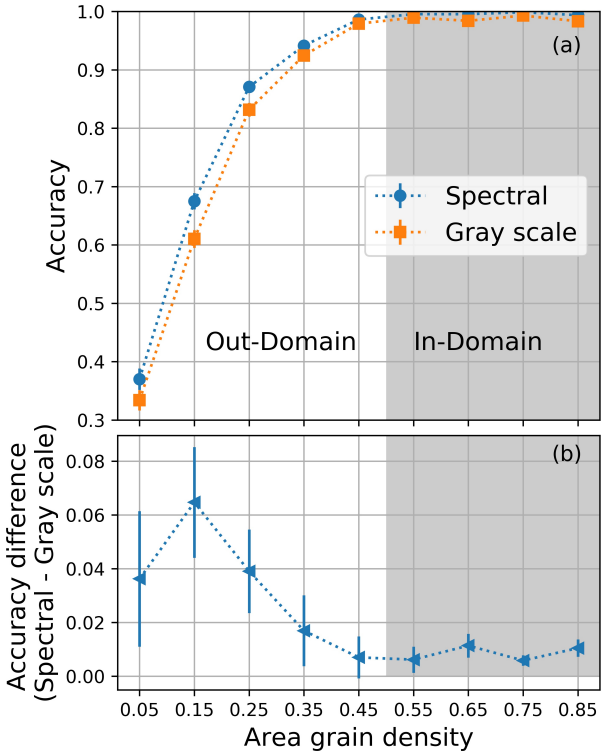


**Figure 10.** Model accuracy as function of kernel area ratio, equation 2, within an evaluated image window from the validation set. (a) shows the accuracy of the ResNet model applied to hyperspectral (Spectral) and grayscale (Gray scale) images. (b) shows the difference in performance between spectral and gray scale. The results are collected in bins of $\rho_P = 0.1$ with points placed at average values. Gray shading (In-Domain) illustrates the kernel area ratio of the windows used for model training.

## Discussion

In this paper, we study potential designs and implementations of deep CNN models capable of bulk grain sample analysis directly from the 3 dimensional HSI data structure. Two different CNN approaches were investigated: 2D ResNet18-spec, a modified version of a ResNet-18 model with 2D convolutional filter with an additional linear spectral down-sampling layer added as input layer and 3D ResNet-18 based on 3D convolutional filters. The results in table 2 show that CNNs which use both spectral and spatial properties of the hyperspectral images improve grain classification compared with classification using either spectral and spatial properties alone. However, using spatial properties alone only reduce the classification accuracy on the test set by 1 %, and the results in figure 4c shows that very little improvement is found by using more than 10 average wavelength channels.

2D ResNet18-spec achieved better performance than the 3D ResNet model when trained on the full 68 bin spectral data with 1/3 training time per epoch. The prolonged training time of 3D ResNet is caused by the 3 times larger CNN kernels, due to the the third dimension (7 times for the first layer). Furthermore the spectral dimension is not reduce at the beginning of the network, but only gradually through the network as the hypercube is down-sampled by using a stride of 2. The 2D ResNet18-spec is constructed to reduce spectral dimension quickly while allowing the network to use the spectral information through its three linear combination of all spectral channels in the first layer. This spectral reduction resembles the use of PCA, where the 3 image-channels left after the layer, seen in figure 8, resembles the PCA score images and weights seen in figure 7, resembles PCA loadings, albeit the different weights of the layer are not orthogonal and is therefore more akin to approaches such as independent component analysis (ICA) [34] and sparse coding [35].

The experimental procedure was designed with a significant difference in moisture content between validation and training sets as documented in figure 1. This design ensured that moisture content did not correlate with grain variety classes. Consequently, the highest classification accuracy of the SVM model was found when removing parts of the spectrum related to OH and CH vibrations. This indicates that the spectral classification with SVM is based on spectral features related to other chemical features possibly linked to protein, and scattering effects observed in the kernels which originates from the physical appearance of the grains. Furthermore, the results show that the experimental design has ensured that average kernel moisture content can not be used to improve the varieties classification.

Figure 4 shows that the most accurate grain classification was obtained using a channel-wise standardization which is commonly used for RGB images. Oppositely, spectral pre-processing conventionally used for NIR-spectra classification leads to deterioration of the classification accuracy. A channel-wise image standardization and spectral pre-processing may however be important when classifying or quantifying parameters that are more dependent on the chemical composition of the grains than the physical appearance, unlike variety classification.

Table 3 shows comparison of the classification accuracy as a function of the data acquisition sequence for the spectral SVM model, spatial ResNet18 model applied to gray scale images, and the spatio-spectral ResNet18 model applied to spectral images. The table clearly shows that although the ResNet model applied to spectral images is affected by the change in spectral properties caused by image-setup variation, it is significantly less affected than the spectral SVM model and the 2D ResNet model applied to gray-scale images. The results indicate that deep learning models used directly on the hypercubes are more resilient to experimental variation than spectral or spatial classification methods alone. For sample presentation, a significant effect on accuracy of varying kernel area ratio between training and validation images was demonstrated in figure 10. The results showed that gray-scale image classification was more affected by the change in kernel area ratio than

the hyperspectral classification, implying that adding the spectral dimension improved the models resilience toward sample presentation variations.

The accuracy on the spectral classification may have been improved by using neural network models. Zhou et al[10] demonstrated that a one dimensional CNN could classify 30 wheat varieties with 93 % accuracy on an independent test set based on single kernel spectra obtained from hyperspectral images. Zhou et al also showed that using SVM or PLS-DA models for classifying the varieties result in significant lower accuracy. It is noted that their SVM model revealed an accuracy of 86 %, which is similar to 86.5 % accuracy observed on the test dataset in this paper. Using 1D CNN for spectral classification is therefore unlikely to change the general conclusion of this paper, as the 93 % classification accuracy on 30 classes reported by Zhou et al is lower than the $> 99$ % accuracy obtained with the gray-scale or hyperspectral CNN approaches shown in this paper.

With gray-scale imaging and sensor fusion, similar results has been reported as the one presented in this paper. Pourreza et al. [16] showed that it was possible to obtain an 98.15 % classification accuracy of nine wheat varieties from visual gray scale images using a combination of linear discriminate analysis and conventional image textual feature extraction. Similarly, Özkan et al. [19] reported a classification accuracy on 40 variety wheat samples classification between 99-100 % using sensor fusion between colour image from visible light (RGB) and short wave infrared and visible-near infrared gray scale images. These results together with the results documented in this paper, highlights that grain classification can be done well using spatial dimension alone.

It should be noted that previous results reported can not be directly compared as sample types, sample presentation, experimental protocol, and image setups, etc are different and are heavily influencing the obtained results. As the scope of this research was to examine possible methods for applying deep CNN directly to the full 3 dimensional HSI datastructure, a simple dataset was used which contained relatively few different grain varieties with each variety harvested from the same field. Thereby, the biological variations caused by different harvest place or year is not included and the limited number of grain varieties limits the complexity of classification. At this point, a bulk sample image library does not exist for the cereal classification problem, which makes cross model comparison subject to sample set variation and complicates model testing as large datasets needs to be constructed for each new research project. To properly test methods for classifying kernels with HSI and to make a better comparison of methods for bulk classification between research projects, a large and biological diverse grain kernel HSI dataset is needed similar to what is known from remote sensing[8]. Such a dataset would need to include different harvest locations and years as well as a large number of different classes to properly test the performance of new models.

Albeit, the number of wheat variety used for the experiments in this paper is relatively low, the findings show how deep learning CNN models can be used for grain classification with high accuracy without performing individual spatial and/or spectral feature selection. Thereby, the results show a potential strategy for including both spatial and spectral information directly using CNN. As some of the used grain varieties are very similar, the results presented in this paper further indicates that the method may be used not only to differentiate grain varieties, but may even be used to identifying grains belong to the same grain sample, which could have potential in screening for adulteration of high value grains with similar but lower value variants. The use of a spectral-down sampling layer has the advantage that its weights can be used to identify the most important wavelengths for predicting correct classes. This knowledge may enable problem specific design of faster multi-spectral camera setups with wavelengths sensitivity selected based on the most important wavelengths according to the spectral down-sampling layer. A fast multi-spectral camera could allow for fast in-line screening of grains.

The finding presented in this paper is based on the variety classification. However, the method of using spatio-spectral information with CNN architectures adapted to hyperspectral images of bulk cereal samples has the potential to find use in other cereal assessment applications where single seed evaluation is not needed or impossible, and could potentially be used for in-line monitoring. For problems where non-linear relationship in the spectral domain are expected to be important, the linear spectral down-sampling layer in 2D ResNet18-spec, could be adjusted by adding non-linear activation functions thereby allowing the model to learn a non-linear mapping of the data into a low dimensional representation.

## Conclusion

In this paper, it is demonstrated how deep CNN models conventionally used for RGB-image classification, such as ResNet, can be employed for bulk cereal hyperspectral image analysis using both spectral and spatial features. The results show that utilizing both spectral and spatial properties in combination through training a Resnet-18 model with a linear spectral reduction layer can classify 8 different grains (6 wheat varieties, 1 spelt wheat, and one rye) with $99.75 \pm 0.02$ % accuracy on a separate test set. This is a significant improvement compared with the $86.5 \pm 0.1$ % test accuracy obtained using only spectral information and non-linear SVM models and $98.70 \pm 0.01$% accuracy when training ResNet models on an average gray scale NIR image. Both, spatial, spectral, and the combined spatio-spectral models presented in the paper are largely affected by sample presentation and imaging setup variations. The results indicate that for classification bulk cereal samples, the combination of hyperspectral images and deep CNNs, can outperform purely spectral or spatial methods in terms of robustness as well as classification accuracy.

However, the study finds that for variety classification, the accuracy improves significantly using just a few spectral channels, and using a resolution of more than 10 channels yields little or no extra accuracy. The ResNet model shows the highest classification accuracy when used on raw data without applying spectral correction algorithms conventionally used within NIR-spectroscopy. However, a channel-wise standardization, similar to what is used for RGB image classification was found to improve the ResNet model classification accuracy significantly. These

result in combination with results based on purely spectral classification show that the grains spatial properties are significantly more important than their spectral for variety classification, in agreement with previous studies.

## Declaration of conflicting interests

The Authors declare that there is no conflict of interest

## Data availability statement

The hyperspectral image data and reference parameters presented in this manuscript is available at https://doi.org/10.17894/ucph.f8c7feeb-3b27-4bd2-ba6d-6d44a4ab4330. This paper should be cited if the dataset is used in a scientific context or publication.

## References

[1] Feng L, Zhu S, Liu F et al. Hyperspectral imaging for seed quality and safety inspection: a review. *Plant Methods* 2019; 15(1). DOI:10.1186/s13007-019-0476-y.

[2] Femenias A, Gatius F, Ramos AJ et al. Standardisation of near infrared hyperspectral imaging for quantification and classification of DON contaminated wheat samples. *Food Control* 2020; 111: 107074. DOI:10.1016/j.foodcont.2019.107074.

[3] Barbedo JG, Guarienti EM and Tibola CS. Detection of sprout damage in wheat kernels using NIR hyperspectral imaging. *Biosystems Engineering* 2018; 175: 124–132. DOI:10.1016/j.biosystemseng.2018.09.012.

[4] Bao Y, Mi C, Wu N et al. Rapid classification of wheat grain varieties using hyperspectral imaging and chemometrics. *Applied Sciences* 2019; 9(19): 4119. DOI:10.3390/app9194119.

[5] Caporaso N, Whitworth MB and Fisk ID. Protein content prediction in single wheat kernels using hyperspectral imaging. *Food Chemistry* 2017; 240: 32–42. DOI:10.1016/j.foodchem.2017.07.048.

[6] Caporaso N, Whitworth MB and Fisk ID. Application of calibrations to hyperspectral images of food grains: example for wheat falling number. *J Spectral Imaging* 2017; 6(a4). DOI:https://doi.org/10.1255/jsi.2017.a4.

[7] Caporaso N, Whitworth MB and Fisk ID. Near-infrared spectroscopy and hyperspectral imaging for non-destructive quality assessment of cereal grains. *Applied Spectroscopy Reviews* 2018; 53(8): 667–687. DOI:10.1080/05704928.2018.1425214.

[8] Signoroni A, Savardi M, Baronio A et al. Deep learning meets hyperspectral image analysis: A multidisciplinary review. *Journal of Imaging* 2019; 5(5): 52. DOI:10.3390/jimaging5050052.

[9] Qiu Z, Chen J, Zhao Y et al. Variety identification of single rice seed using hyperspectral imaging combined with convolutional neural network. *Applied Sciences* 2018; 8(2): 212. DOI:10.3390/app8020212.

[10] Zhou L, Zhang C, Taha MF et al. Wheat kernel variety identification based on a large near-infrared spectral dataset and a novel deep learning-based feature selection method. *Frontiers in Plant Science* 2020; 11. DOI:10.3389/fpls.2020.575810.

[11] Weng S, Tang P, Yuan H et al. Hyperspectral imaging for accurate determination of rice variety using a deep learning network with multi-feature fusion. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 2020; 234: 118237. DOI:10.1016/j.saa.2020.118237.

[12] Zhang J, Yang Y, Feng X et al. Identification of bacterial blight resistant rice seeds using terahertz imaging and hyperspectral imaging combined with convolutional neural network. *Frontiers in Plant Science* 2020; 11. DOI:10.3389/fpls.2020.00821.

[13] Zhu, Zhou, Zhang et al. Identification of soybean varieties using hyperspectral imaging coupled with convolutional neural network. *Sensors* 2019; 19(19): 4065. DOI:10.3390/s19194065.

[14] Laidig F, Piepho HP, Rentel D et al. Breeding progress, environmental variation and correlation of winter wheat yield and quality traits in german official variety trials and on-farm during 1983–2014. *Theoretical and Applied Genetics* 2016; 130(1): 223–245. DOI:10.1007/s00122-016-2810-3.

[15] Zapotoczny P and Ropelewska E. Application of hyperspectral imaging for cultivar discrimination of malting barley grains. *Agricultural Engineering* 2016; 20(3): 207–217. DOI:10.1515/agriceng-2016-0058.

[16] Pourreza A, Pourreza H, Abbaspour-Fard MH et al. Identification of nine iranian wheat seed varieties by textural analysis with image processing. *Computers and Electronics in Agriculture* 2012; 83: 102–108. DOI:10.1016/j.compag.2012.02.005.

[17] Choudhary R, Paliwal J and Jayas D. Classification of cereal grains using wavelet, morphological, colour, and textural features of non-touching kernel images. *Biosystems Engineering* 2008; 99(3): 330–337. DOI:10.1016/j.biosystemseng.2007.11.013.

[18] Pang L, Men S, Yan L et al. Rapid vitality estimation and prediction of corn seeds based on spectra and images using deep learning and hyperspectral imaging techniques. *IEEE Access* 2020; 8: 123026–123036. DOI:10.1109/access.2020.3006495.

[19] Özkan K, Işık Ş and Yavuz BT. Identification of wheat kernels by fusion of RGB, SWIR, and VNIR samples. *Journal of the Science of Food and Agriculture* 2019; 99(11): 4977–4984. DOI:10.1002/jsfa.9732.

[20] He K, Zhang X, Ren S et al. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 770–778.

[21] Li Y, Zhang H and Shen Q. Spectral–spatial classification of hyperspectral imagery with 3d convolutional neural network. *Remote Sensing* 2017; 9(1): 67. DOI: 10.3390/rs9010067.

[22] Zapotoczny P. Discrimination of wheat grain varieties using image analysis: morphological features. *European Food Research and Technology* 2011; 233(5): 769–779. DOI:10.1007/s00217-011-1573-y.

[23] Ioffe S and Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*.

[24] Nair V and Hinton GE. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*.

[25] Hara K, Kataoka H and Satoh Y. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[26] He K, Zhang X, Ren S et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *pre-print* 2015; arXiv:1502.01852.

[27] Smith LN. Cyclical learning rates for training neural networks. *pre-print* 2017; arXiv:1506.01186v6.

[28] Rinnan Å, van den Berg F and Engelsen SB. Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends in Analytical Chemistry* 2009; 28(10): 1201–1222. DOI:10.1016/j.trac.2009.07.007.

[29] Savitzky A and Golay MJE. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry* 1964; 36(8): 1627–1639. DOI:https://doi.org/10.1021/ac60214a047.

[30] Martens H, Jensen SA and Geladi P. Multivariate linearity transformations for near infrared reflectance spectroscopy. In *Nordic Symp. on Applied Statistics, Stavanger, Norway*.

[31] Barnes RJ, Dhanoa MS and Lister SJ. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied Spectroscopy* 1989; 43(5): 772–777. DOI:10.1366/0003702894202201.

[32] Otsu N. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics* 1979; 9(1): 62–66. DOI:10.1109/tsmc.1979.4310076.

[33] Yuille AL and Liu C. Deep nets: What have they ever done for vision? *International Journal of Computer Vision* 2021; 129: 781–802. DOI:10.1007/s11263-020-01405-z.

[34] Hyvärinen A and Oja E. Independent component analysis: Algorithms and applications. *Neural Networks* 2000; 13(4-5): 411–430.

[35] Olshausen BA and Field DJ. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 1996; 381: 607–609.