Robust classification approach for segmentation of blood defects in cod fillets based on deep convolutional neural networks and support vector machines and calculation of gripper vectors for robotic processing

Ekrem Misimi[a*], Elling Ruud Øye[a], Øystein Sture[a], John Reidar Mathiassen[a]

[a] *SINTEF Fisheries and Aquaculture, NO-7465 Trondheim, Norway*

** Corresponding author: ekrem.misimi@sintef.no; Tel.: +47 982 22 467*

ABSTRACT

Despite advances in computer vision and segmentation techniques, the segmentation of food defects such as blood spots, exhibiting a high degree of randomness and biological variation in size and coloration degree, has proven to be extremely challenging and it is not successfully resolved. Therefore, in this paper, we propose an approach for robust automated pixel-wise classification for segmentation of blood spots, focusing specifically on challenging texture-uniform cod fish fillets. A multimodal vision system, described in this paper, enables perfectly aligned RGB and D-depth images for localization of segmented blood spots in 3D. Classification models based on 1) Convolutional Neural Networks - CNN and 2) Support Vector Machines - SVM for the classification of defective fillets were developed. A colour-based, pixel-wise and SVM-based model was developed for accurate segmentation and localisation of blood spots resulting in 96% overall accuracy when tested on whole fillet images. Classification between normal and defective fillets based on GPU (Graphical Processing Unit) -accelerated CNN classification model achieved 100% accuracy, versus the SVM-based model achieving 99%. We present a novel data augmentation approach that desensitizes the CNN towards shape features and makes the CNN to focus more on colour. We show how pixel-wise classification

is used for an accurate localization of blood spots in 3D space and calculation of resulting 3D gripper vectors, as an input to robotic processing.

Keywords: Image Segmentation; RGB-D image; Robotics; Support Vector Machines; Deep Convolutional Neural Networks; Data Augmentation.

1. Introduction

Blood spots and discolouration resulting from inappropriate bleeding are detrimental to fillet flesh quality [1]. The visual effect of residual blood in fillets reduces consumer acceptance and the market value of the product. Currently, fillets with blood spots are manually sorted and trimmed to remove parts that are discoloured due to the presence of blood. The industry requires a robust, rapid, non-invasive and cost-efficient method for the effective discrimination of normal and defective fillets, which automatically segments and localises blood spots using image technologies. Blood spot segmentation is a scientific challenge that remains unresolved despite recent developments in image-based segmentation techniques. Image-based segmentation continues to be a very challenging problem, and is highly application dependent [2]. The segmentation of blood spots in fillet muscle tissue falls into the category of hard-to-solve challenges due to high levels of randomness, high variation in colour, spectral similarity of blood spots with other similar defects and inherent biological variation encountered in biological raw materials. For this reason, addressing this challenge with a cost-efficient multimodal imaging system has great value, both generically and in terms of practical application. While recent imaging techniques, combined with machine learning [3-5, 6, 7], have been shown to be efficient tools for food quality assurance, it is also shown that food applications and recognition are challenging topics in computer vision [8].

Blood spot detection and segmentation in raw material has proved to be very challenging to

automate. Mertens et al. [9] performed a spectral characterisation of egg shells to detect blood spots and concluded that brown pigments and other discolouration of the shells interfere with the peak detection of blood (at 577 nm) and thus makes the detection of blood spots challenging. Balaban et al. [5] developed an image analysis method to quantify gaping and bruising, and the presence of blood spots, observed on salmon fillets, by adaptively applying an *L* (Lightness) threshold value. The authors suggested that a robust blood spot detection system was required, based on a specifically tailored classification algorithm. Although popular due to their simplicity [10], image thresholding segmentation methods using traditional histogram-based thresholding cannot separate areas exhibiting high similarity in grey scales not belonging to same regions.

Image segmentation still remains an important area of research in the field of computer vision [2], and several approaches and methods have been proposed to solve this generic problem. These approaches are categorised according to methodology: histogram thresholding methods, clustering methods, edge detection, region methods and graph methods [2, 10 and 11]. For biological raw materials, image segmentation approaches are extremely application dependent. Image structure and information exhibit high levels of variation and randomness, making the segmentation operation even more challenging. Sometime, as in the case of cod fish fillets, the high degree of uniformity of texture of the fillet muscle image is a disadvantage and disabling factor in, for example, including texture alongside the colour as features to be used for development of robust segmentation approaches.

A prerequisite for an effective automated system is that following trained learning, it must be able to classify objects into respective class categories based on features detected on images. The selection of the appropriate classification algorithm is, therefore, key to this process. The most commonly used classification approaches for generic non-food and food related applications are a) statistics-based, and b) those based on Neural Networks (NN). In recent

years, Support Vector Machines (SVMs) [12] have emerged as powerful classification algorithms for food applications due to their excellent performance in a variety of quality inspection tasks [13] as it can be used to solve both classification and regression problems. SVM classification algorithms has already been successfully used in several food applications such as prediction of product quality in industrial bakery processes, prediction of beef tenderness using image colour and texture features [14, 15]. Du and Sun [16] used low dimensional colour features and support vector machine algorithm to perform an automated classification of pizza sauce spread achieving 96.6% classification accuracy on the test set. The concept of deep learning is also emerging as a powerful machine learning method that allows computational models composed of multiple processing layers to learn representations of data containing multiple levels of abstraction and has dramatically improved the state-of-the-art of visual recognition applications [17]. Kagaya et al. [18] used a convolutional neural network for recognizing food images and they observed that the network achieved significantly better performance accuracy (93.8 %) than the baseline method (89.7%). In deep learning, data augmentation [19] is important since in practice the amount of available data for training the network is limited. Therefore, data augmentation procedure must be performed correctly so that transformations performed in the image does not change the image class.

3D image information is valuable in applications involving robotic processing of food and calculation of respective gripper vectors, containing the pose information for the gripper, is necessary in such applications [20]. Misimi et al. [20] demonstrate how 3D information from the Kinect v2 RGB-D camera is used to calculate the correct grasping point for 3D vision based robotic harvesting of chicken fillets.

The main research objectives of this study were: **a)** to develop a robust, colour-based pixel-wise classification algorithm for blood spot segmentation in fillets as an example of objects with high intra-class variance when it comes to size, colour and localization of blood spots, **b)**

to develop a model for accurate classification of normal and defective fillets, **c)** to develop an approach for perfectly aligned RGB-D images that can make use of pixel-wise classification for accurate localization of blood spots in 3D and calculation of gripper vectors for robotic processing**; d)** to acquire a deeper understanding and visualisation of how changes in SVM hyperparameters influence pixel-wise classification in general and blood segmentation in particular, and **e)** to exploit the capabilities and acquire deeper understanding of CNN for classification of cod fillets as an example of food objects and appropriateness of current data augmentation techniques for such applications.

To the best of our knowledge, no work has been published on the automated segmentation of blood spots or similar defects in food objects based on perfectly per-pixel aligned RGB-D images and robust pixel-wise classification and localisation in 3D space. Contribution on visualizing the effects of change of SVM parameters in resulting classification and segmentation accuracy is also original. This paper investigates the application of CNN-based deep learning classification in food sorting applications. For this reason, the knowledge obtained by means of this study on the use and understanding of deep learning for raw food material classification is original. The data augmentation approach used to reduce the sensitivity of the CNN approach in terms of shape and increased colour sensitivity is also novel.

The rest of the paper is organized as follows: in materials and methods section we describe the collected datasets, multimodal vision system overview, and the approach for classification and segmentation of blood spots. In results and discussion section, we show in detail our results and discussion regarding the CNN and SVM classifications model, we visualize and discuss the effect of SVM hyperparameters in actual pixel-wise classification and segmentation of blood spots and we calculate the 3D gripper vectors for robotic processing. In future work section are given some solid future research directions, and finally in conclusion section, we draw some final conclusions.

2. Materials and methods

2.1. Sample preparation: Fish fillets taken from farmed Atlantic cod (Gadus morhua) were differentiated by a qualified human inspector into two categories: a) normal (n=33), and b) defective (d=32), with mean length 48.5 cm ± 5.8 cm. They were subsequently shipped from the Norway Seafoods (Melbu, Norway) fish processing company to SINTEF SeaLab in Trondheim where they were stored at 4^{o}C prior to imaging.

2.2. Computer vision system used to acquire the image dataset

Currently existing vision cameras such as 3D SICK IVP ColorRanger, or RGB-D Kinect v2 don't generate aligned RGB and D-depth images. This is very often a drawback when combination of both RGB and D-depth information is needed for accurate localization of regions of interest and defects in 3D space for robotic applications [21]. The multimodal vision system in this paper consisted of a colour imaging line scan CMOS camera, Grasshopper 3 (GS3-U3-23S6CC, Point Grey, Canada), with a USB 3.0 interface, enabling aligned RGB and 3D images. The Region of Interest (ROI) used for imaging the fillets was 1376 x 64, with an exposure time of 500µs. The working distance to the camera was 54 cm and the tilt angle was 17 degrees. Each fillet was placed on a conveyer belt for image acquisition. A laser emitting a 100 mW red uniform laser line at 660 nm wavelength, with a fan angle of 30 degrees, was used in triangulation mode to acquire 3D and reflectance images of the cod fillets. White illumination used to acquire RGB images was provided by a flexible white LED strip, with colour temperature of 4000K and colour rendering index RI larger than 75. To enable simultaneous acquisition of RGB and 3D fillet images using the same camera, the LEDs strips and the laser were triggered alternately for every other frame. The resulting RGB and 3D images used for developing of classification models for segmentation of blood spots had a 2650x1317 resolution.

2.3. Image pre-processing and feature extraction

2.3.1 Colour calibration of the RGB images*:* A Gretag Macbeth colour checker with 24 patches (from Color-Science AG, Hinwil in Switzerland) was used to perform subsequent colour calibration of images in RGB space using the provided reference sRGB values (from X-Rite, Munich in Germany). The colour correction matrix was calculated by finding the least squares solution that minimises the error between the mean of the measured RGB values of each patch, and the corresponding reference sRGB value. The 4 x 4 colour correction matrix A was found by calculating;

$$\min_{A} \|AC - C^*\| \tag{1}$$

where C is a matrix containing the measured RGB mean values for all the 24 colour checker patches

$$C = \begin{bmatrix} R_1 & R_2 & \cdots & R_{24} \\ G_1 & G_2 & \cdots & G_{24} \\ B_1 & B_2 & \cdots & B_{24} \\ 1 & 1 & \cdots & 1 \end{bmatrix} \tag{2}$$

and C* is a matrix containing the correct reference RGB-values for the corresponding patches

$$C^* = \begin{bmatrix} R_1^* & R_2^* & \cdots & R_{24}^* \\ G_1^* & G_2^* & \cdots & G_{24}^* \\ B_1^* & B_2^* & \cdots & B_{24}^* \\ 1 & 1 & \cdots & 1 \end{bmatrix} \tag{3}$$

2.3.2 Pre-processing, segmentation and colour spaces

Figure 1 shows a flowchart of the processing operations that resulted in the images that were used for feature extraction and training of the classification models. The 3D image was used to generate a binary mask which in turn was used with the colour-calibrated image to segment the cod fillets from their background. The RGB fillet images were converted to CIELab colour space by first converting the RGB images to an XYZ matrix system according to the following equation [22]:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.4124 & 0.3576 & 0.1805 \\ 0.2126 & 0.7152 & 0.0722 \\ 0.0193 & 0.1193 & 0.9505 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{4}$$

and then by calculating the L, a and b values, resulting in L, a, and b image channels, as shown in Figure 2. Conversion between RGB and HSV colour spaces was performed according to the expressions found in [22] (Fig. 2). The ΔE image (Figs. 1 and 2) in the CIELab colour space was calculated for every pixel in the image as the value of the difference between that pixel's Lab colour and the average Lab value for the entire fillet according to the formula:

$$\Delta E = \sqrt{(L_m - L_i)^2 + (a_m - a_i)^2 + (b_m - b_i)^2} \quad (5)$$

where L_m, a_m, b_m are mean values for the entire fillet, while L_i, a_i, b_i are values for each i-pixel of the fillet image.

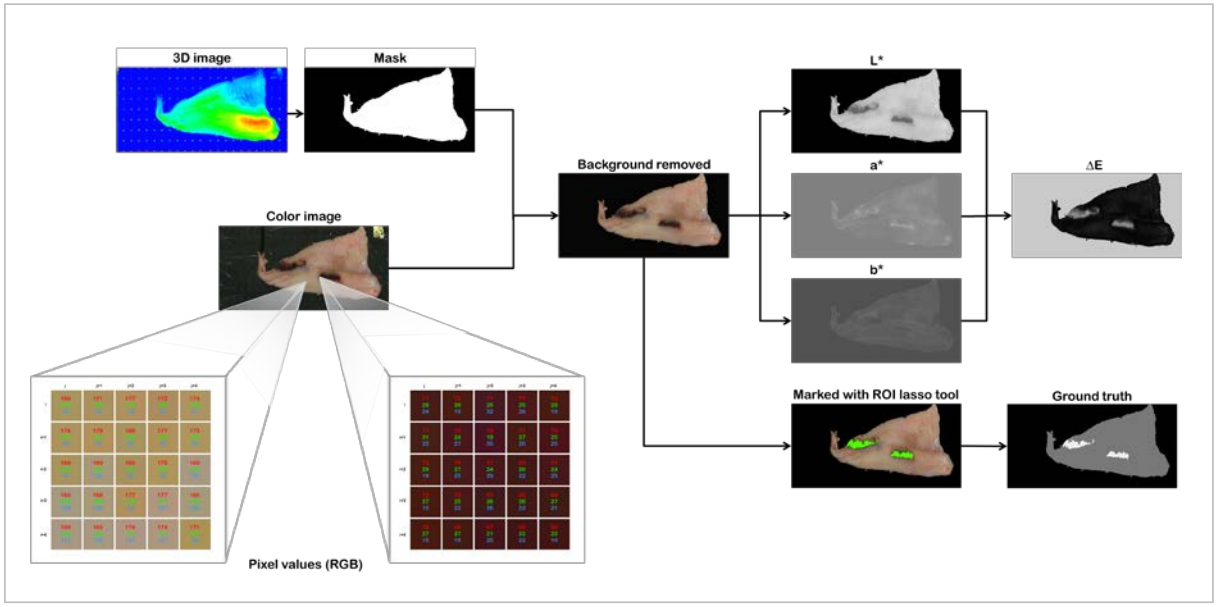


Figure 1. RGB and 3D images of an example fillet and the sequence of computer vision operations used to generate images, features, and the ground truth used for training of the classification algorithms. RGB pixel values of the normal muscle are higher (lighter colour) than the pixel values of the blood spots (dark colour).

2.3.3 Ground truth

In order to facilitate supervised learning of the pixel-wise classification models for blood spot segmentation, a set of ground truth images were generated by manual labelling the blood spot regions according to the following procedure: 1) A trained human inspector had previously provided input as to what constituted a blood spot in the form of labels attached to each fillet; 2) Prior to imaging in the lab, we noted for each fillet the localisation and size of the blood

spot(s); 3) All this information was available to the researcher who performed manually the final ground truth labelling on images, marking manually the boundaries of the blood spots. In this way, a two-level validation of blood spots (steps 1 and 2 above) was completed prior to establishment of the final ground truth. As a result, all images were manually labelled and subdivided into regions belonging to three categories (Fig. 1): 1) background (black), 2) normal fillet muscle tissue (grey) and 3) blood spots (white). Manual labelling as a means of facilitating supervised learning is a well-known technique used in the development of automated classification models [10].

2.3.4 Hand-engineered feature extraction for SVM classification

Feature extraction is a key success criterion during the design of a pattern recognition system. It requires that features are extracted that exhibit the most distinctive characteristics from among different classes [8]. Hand-engineered features are traditionally engineered features (hand-crafted), which are used in training the traditional classifiers as opposed to learned features in deep learning which are automatically learned by the network [17]. Feature extraction was carried out with a view to performing a two-fold classification involving a) discrimination between normal and defective fillets, and b) a pixel-wise classification for blood spot segmentation and localisation. We selected features that were not only of direct relevance to the application but also rapidly computable. Table 1 shows a complete list of the features that were extracted for this study. In the case of discriminating between normal and defective fillets, so-called FC_m features represent mean colour values extracted from the fillet image, while for blood spot detection and localisation, the pixel-level features FC_i shown in Table 1

were extracted for each pixel of the image in question (Fig. 2).

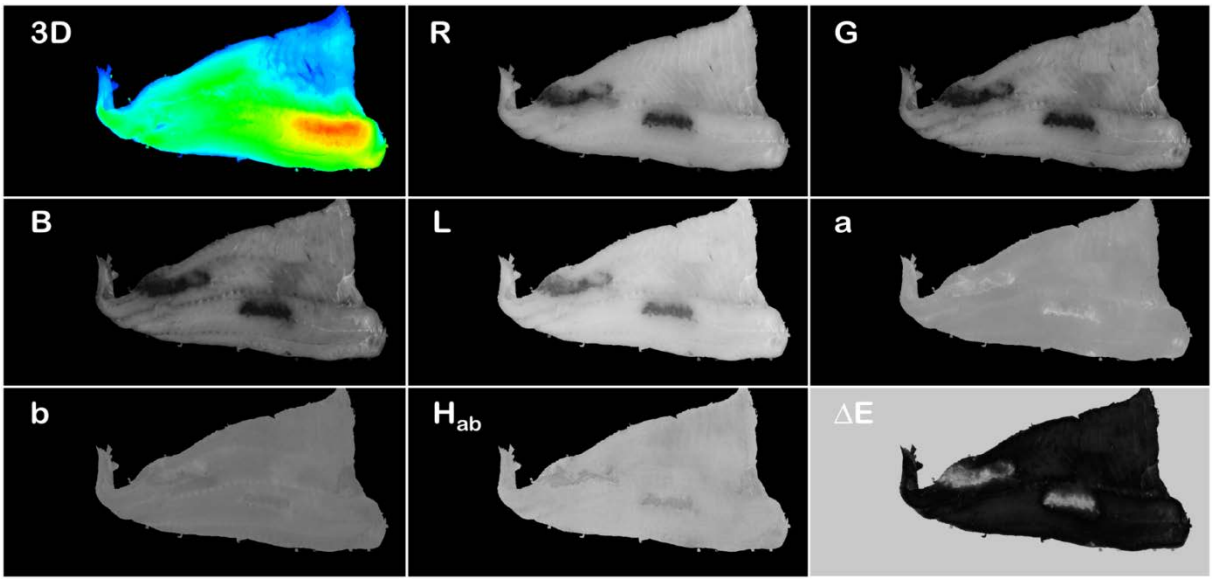


Figure 2. Fillet images displayed in different colour spaces used for the extraction of features for discrimination and blood segmentation and localisation in 3D space. 3D – 3D image, R-Red channel of RGB, G-Green channel of RGB, B-Blue channel of RGB, L-lightness channel of Lab, a-redness channel of Lab, b-yellowness channel of Lab, Hab-Hue channel of HSV, ΔE-Delta E.

Extraction of the R, G, B, L, a, b, Hue, Chroma parameters and ΔE colour means (Table 1) and pixel-level features is carried out according to the methods and formulae reported in [23], while Whiteness was defined as $W = L - 3b$. The mean value FC_m for a single colour plane was defined by

$$FC_m = \frac{1}{|S|}\sum_{ij \in S} C_{ij} \quad (6)$$

where C is the MxN image matrix (M-rows, N-columns) containing all pixel values of a single colour plane (channel) in the particular colour space, and S is the set of index pairs (i, j) for all pixels covering only the fillet image. As is seen in Figures 1 and 2, an S-set of index pairs was generated from the binary mask and used to segment the fillet from its background.

Table 1. Extraction of the feature set for operation 1) discrimination between normal and defective fillets and 2) the pixel-wise classification for blood segmentation. Feature selection is according to the Fisher's Discriminant Ratio (FDR) criterion.

Feature Set	Normal vs defective	Pixel-wise classification	FDR	Feature Ranking
R-Red	187.5±5.7 169.4±11	*R(i,j)* pixel level feature	2.1586	B
G-Green	158.3±7 132.6±10.1	*G(i,j)* pixel level feature	4.3171	G
B-Blue	133.6±5.4 111.6±7.5	*B(i,j)* pixel level feature	5.6669	L
L-Lightness	83.7±1.3 78.4±2.4	*L(i,j)* pixel level feature	3.8467	W
a-redness	3.25±1 5.8±1.4	*a(i,j)* pixel level feature	2.0667	R
b-yellowness	9.5 ± 0.9^{nsd} 10 ± 1.1^{nsd}	*b(i,j)* pixel level feature	0.1197	a
H-Hue	71.1±6.6 60.1±7.8	*H(i,j)* pixel level feature	1.1975	ΔE
C-Chroma	10.2±0.8 11.7±0.9	*C(i,j)* pixel level feature	1.6860	C
W-Whiteness	55.1±2.7 48.2±2.8	*W(i,j)* pixel level feature	3.0314	H
ΔE-Delta E	5.4±0.3 6.6±0.8	*ΔE (i,j)* pixel level feature	1.8864	b

nsd - not significantly different

2.3.5 Feature selection

Feature selection is a sorting procedure that, for a given set of extracted features, consists of choosing the most important features (to reduce numbers) while at the same time also retaining those that display the maximum amount of discriminatory information [24, 25]. One of the main reasons why feature selection is required is to increase the generalisation properties of the classification model. It has been shown that in industrial applications where the acquisition of large datasets is an expensive business [26, 27], the ratio *T/l* between the available samples in a dataset *T*, and the number of features *l* used to train a model, is directly proportional to the generalisation properties of the model. The higher the *T/l* ratio, the better the generalisation properties of the model [25]. The so-called classifier error estimate improves as this ratio becomes higher and, in [27], it is suggested that this ratio should be as high as 10 to 20 for some applications. The first step of the feature selection procedure is based on a statistical hypothesis testing technique looking into whether there was a significant difference in values

(P<0.05) for the feature in question for the different classes [25]. Subsequently, we ranked these features according to a 'class separability parameter'. In this case, we selected the Fisher's Discriminant Ratio (FDR), which is commonly employed to quantify the discriminatory power of individual features between two classes, and which for a scalar feature *y* in a 2-class classification problem is defined as [25]:

$$FDR = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} \quad (7)$$

where μ_1, μ_2 are the mean values of feature *y*, while σ_1^2, σ_2^2 represent the variances of *y* in the respective classes (normal and blood). In the equation (7), $(\mu_1 - \mu_2)^2$ is the between-class variance, and $\sigma_1^2 + \sigma_2^2$ the within-class variance.

2.3.6 Feature scaling

We performed a Min-Max feature scaling to ensure that all feature values were scaled to a fixed range [0, 1] according to the following expression:

$$X_i = \frac{(X_i - X_{min})(U_x - L_x)}{(X_{max} - X_{min}) + L_x} \quad (8)$$

where L_x, U_x are the lower and upper limits [0, 1], and X_{max} and X_{min} the maximum and minimum feature values respectively. Feature scaling is important since, during the training of an SVM classifier for feature values with different dynamic ranges, larger feature values may exert a bigger influence in the cost function than those with smaller values [25].

2.4. GPU accelerated deep learning

Deep learning is a branch of machine learning that allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction and has in recent years brought dramatic improvements in the visual recognition area [17]. The advent of GPU computing has enabled training in connection with deep learning neural networks to become up to 10 or 20 times faster. In fact, although machine learning and neural networks have been utilised for decades, two relatively recent trends have been required to spark their widespread use – the availability of massive volumes of training data, and the

emergence of powerful and efficient parallel computing tools provided by GPU computing (of NVIDIA, Santa Clara in the USA). Of particular interest for vision-based applications are the so called Convolutional Neural Networks (CNNs), which are inspired by the visual cortex and subsequently tailored for computer vision [28]. CNNs are designed to process data input in the form of multiple arrays, such as RGB colour images containing three two-dimensional arrays (rows and columns) of pixel intensities for Red (R), Green (G), and Blue (B) channels [17]. CNNs can learn a hierarchy of features automatically by convolving the input image with learned filters to build a hierarchy of feature maps in which each map is a rectangular image [17]. In our study, we used the pre-trained AlexNet 12 [29] for the processing of RGB images. The architecture of AlexNet consists of five convolutional layers each followed by a rectified linear unit (ReLU) layer, brightness or contrast normalisation and overlapping pooling. The convolution layer is the core building block of AlexNet and consists of a set of learnable convolution filters. These filters are small in size and are slid across the RGB input image to produce a 2D array feature map representing a particular feature extracted at all locations [28]. For *l*-convolutional layers, the input image or a feature map of the previous layer is convolved using different filters to produce the respective output feature map of the first layer. The ReLU layer consists of rectifier activation functions in the form $r(z) = max(0, z)$, where z is the input to a neuron. The $r(z)$ parameter is actually a ramp function and is currently the most popular activation function for state-of-the-art deep neural networks [29]. It enables faster and more effective training of deep neural architectures on large and complex datasets, and as such is more effective than traditional sigmoid and hyperbolic tangent activation functions [30]. To reduce overfitting, the method referred to in [29] used so-called ‘dropout’ as a regularisation technique. Dropout is a powerful regularisation technique proposed by Hinton et al. [31], and developed based on observations of the human brain. It operates by means of setting individual outputs in a hidden layer to zero with a probability of 0.5. The neurons that are "dropped out"

297 do not contribute to the forward pass and do not participate in backpropagation.

298 ***2.4.1 Data augmentation approach for fine-training of AlexNet***

299 The performance of deep neural networks is greatly improved by having large training datasets,
300 since this is one of the most efficient methods to reduce overfitting of image data. In industrial
301 food processing applications, the acquisition of large datasets is challenging and costly [26],
302 and one of the most inexpensive ways to expand available image datasets is to enlarge them
303 artificially using label-preserving transformations [29] as part of a process known as data
304 augmentation. This allows expansion of the datasets by applying different transformations to
305 the original images. In our study, a set of augmentations were applied repeatedly and at random
306 to each image in order to obtain an artificially larger dataset. Before any of these augmentations
307 were implemented, the background to each fillet image was carefully removed by setting the
308 background pixels to black. The images were then padded out to be square, and then centred
309 according to the centroid point of the non-black pixels. The black border pixels were then set
310 to assume the mean colour of the fillet. This is also carried out in case of transformations that
311 cause border pixels to enter the image (e.g. rotations). The following augmentations were
312 applied to our original dataset: *Flipping:* flipping was applied (Figs. 3a, b, c) according to a
313 binary Bernoulli distribution. The image was flipped conditionally about the horizontal, vertical
314 or both axes. *Scaling:* the image was scaled (Fig. 3d) randomly within the log-uniform range
315 [1/1.6, 1.6]. Each axis was scaled independently. *Rotation:* the image was rotated (Fig. 3e)
316 within the range [0, 360] degrees about its centre using a uniform probability distribution. *Shear*
317 *image:* shear (Fig. 3f) was applied within the range [-30, 30] using a uniform probability
318 distribution. *Translation:* translation (Fig. 3g) was applied within a [-25, 25] pixel range, which
319 is approximately +/- 10% of the image width and height. As with the scaling, each axis was
320 determined individually according to a uniform distribution. *Smoothing:* to smooth out the
321 edges where the fillet meets the background, a Gaussian smoothing filter ($\sigma = 10$) was applied

(Fig. 3h) on each channel separately. This is not an augmentation per-say, but is performed prior to the next augmentation step which consists of adding Gaussian noise to the image. An additional function involved in this operation was to reduce edge detection of the fillets' outer shape contour. *Gaussian noise:* Once the previous augmentations were applied at random, Gaussian noise was added to each channel (Fig. 3i). Gaussian noise has zero-mean, and a variance of 0.5. The noise was scaled according to the principal component directions of the image. The data augmentation described above resulted in a dataset containing 21,775 samples with more than 10,000 samples in each class. All images in the augmented dataset were RGB

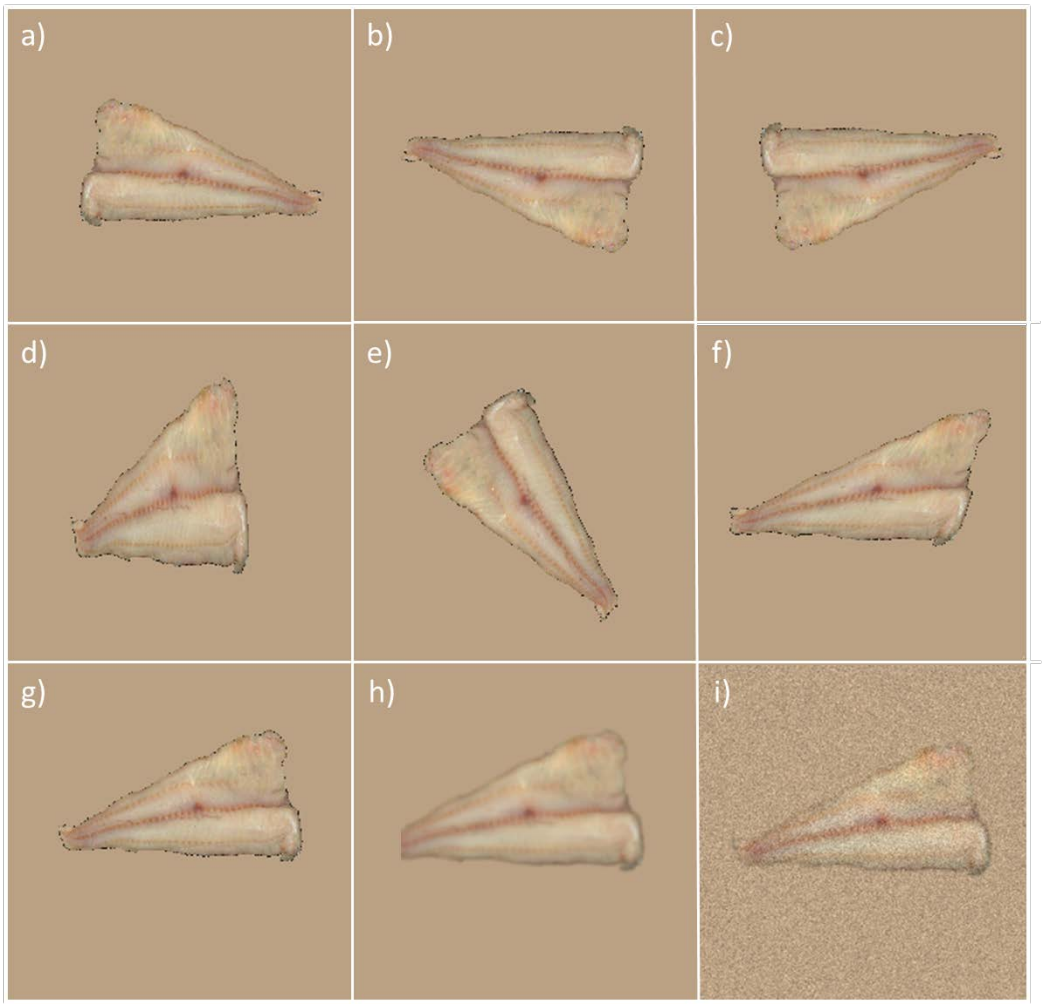


Figure 3. Fillet image transformations as part of data augmentation of the original dataset: a) Fillet image; b) flipped image c) randomly flipped imaged; d) Log-uniform scaled fillet image; e) randomly rotated fillet image; f) shear transform of the fillet image; g) translation of the image; h) image smoothing with a Gaussian filter; i) added Gaussian noise to the fillet image.

images of 256 x 256 pixel resolution, as required by the AlexNet deep learning model. The dataset was labelled to enable supervised training. In fact, a total of 15,327 images were used

for training while approximately 25% (5,108) were retained for validation. The remaining 1,340 images (670 in each class) were set aside for testing. The training based on our datasets was implemented in NVIDIA's DIGITS 2.2 platform for deep learning (from Nvidia, Santa Clara in the USA) on a PC running on Ubuntu, a debian-based Linux operating system. DIGITS 2.2 integrates the Caffe deep learning framework (developed by UC Berkley, USA) and supports GPU acceleration using the CuDNN library to massively reduce training time. The CuDNN library is a collection of GPU-accelerated primitives designed for the CNNs and facilitates implementation of CNN routines such as convolution, pooling, softmax, and neuron activations (sigmoid, ReLU, and tanh). DIGITS 2.2 includes automatic multi-GPU scaling, and in our study we used a GeFORCE GTX 780 Titan GPU processor (from Nvidia, Santa Clara in the USA) with a 3 GB memory. AlexNet was trained for 30 epochs. In our study, fine-training of the AlexNet, generation of the classification model and validation for a dataset containing 21775 images, took approximately one hour.

2.5. SVM model selection and training for classification and pixel-wise segmentation

Support vector machines (SVMs) represent powerful classification algorithms that are relatively insensitive to dimensionality, exhibit excellent performance in general terms, and are suitable for working with high dimensional data [25]. SVM have emerged as powerful classification algorithms for various applications due to their excellent performance in a variety of quality inspection tasks [12, 13]. Since there is no prior knowledge of which SVM parameters are best suited to a given application, it is necessary to perform a parameter search to select the optimal model. C-SVM is a SVM modality that uses C as a regularisation parameter, and Radial Basis Function (RBF) kernel is the first choice due to the ability to nonlinearly map the samples into a higher dimensional space [32]. When using a C-SVM with a RBF kernel, there are two parameters that require tuning – the penalty parameter C, and the free kernel parameter γ. Several methods can be used to perform the parameter search [33]. This procedure is usually

performed by dividing the labelled training set D at random into K folds, i.e. K-disjoint sets of equal size (n/m), where n is the total number of samples used for training and m is the total number of samples in the validation set. In K-fold cross-validation, one of the folds is used as a validation set, while the remaining K-1 folds are used for training and the holdout approach repeated K times (folds) [34, 35]. The benefits of using K-fold cross-validation are that the validation folds are independent [36] which minimises the impact of data dependency [37]. The leave-one-out (LOO) cross-validation [34] takes this to the extreme by validating a single sample at a time until every sample has been used. LOO is often computationally expensive and is thus rarely used in practice. In this study, a 10-fold stratified cross-validation approach was employed. The folds were stratified so as to contain approximately the same proportions of labels as the original dataset. A stratified cross-validation also ensures that each class is equally distributed across all random splits. In this paper, a split of 80 training and 20 validation sets has been applied.

In terms of parameter selection for the SVM algorithm, it is possible to perform an exhaustive search by attempting all parameter combinations within a certain region with a defined spacing. This approach is known as a grid search algorithm [38]. Other, more advanced methods of hyperparameter optimisation exist in which the search is directed intelligently through sequential steps based on randomly chosen trials [39]. Since SVM algorithms exhibit a low number of hyperparameters, an exhaustive search is therefore perfectly feasible from a computational standpoint, especially if the search is performed in parallel [33]. We selected a grid search strategy for the RBF kernel function which involved selecting the optimal parameter pairs within the following ranges: C= $[2^{-25} \dots 22^{30}]$ and $\gamma = [2^{-25} \dots 22^{30}]$[33]. The exhaustive grid search used in this paper was evaluated at a fine-grained exponential scale 2^x with a step size of 1. We also performed a wide exhaustive grid search in the range $C = [2^{-150} \dots 12^{150}]$ and $\gamma = [2^{-75} \dots 2^{30}]$ on an exponential scale of 2^x with a step size of 5. The

parameter pair (C, γ) that maximised the prediction rate after completion of a 10-fold stratified cross-validation was chosen and used to train a model using the full training set, as recommended in [33]. This model was then used to evaluate the final performance of the validation set. All image processing steps, training, and validation were performed using a desktop PC with an Intel-Core i7-4770K processor (from Intel, Santa Clara in the USA) and an 8MB cache and processor speed of 3.9 GHz, 16 GB RAM (DDR3), and GTX 780 (from NVIDIA, Santa Clara in the USA) GPU processor with 3GB RAM.

2.6. Performance evaluation

The performance metric used for evaluation of the classification algorithms was overall accuracy, while for the pixel-wise classification we also used the parameters True Positive rate - TPR, True Negative rate-TNR, as well as the CPU time used to segment a single fillet image and the segmentation error rate - ER [10] to evaluate the performance of blood segmentation. ER is defined as the ratio between the mis-classified image pixels over the total image pixels:

$$ER = \frac{N_f + N_m}{N_t} \text{x} 100\% \quad (13)$$

where N_f is the number of false-detection image pixels, N_m is the number of miss-detection image pixels, and N_t is the total number of image pixels [10].

3. Results and Discussion

3.1. Computer vision and feature selection

The flowchart of image pre-processing steps for fillet images for classification and feature extraction is shown in Figure 1, with the resulting images in Figure 2. A distinct colour feature set was built up based on feature extraction consisting of a) average fillet colour values in several colour spaces – $FColour = \{R_i,\ G_i,\ B_i, L_i, a_i,\ b_i, H_i, C_i,\ W_i, \Delta E_i\}$, where *i* is the fillet sample number in the dataset; and b) pixel-level colour features (Figure 2) $P_{(ij)} = (R_{ij}, G_{ij}, B_{ij}, L_{ij}, a_{ij}, b_{ij}, H_{ij}, C_{ij}, W_{ij}, \Delta E_{ij})$, for a pixel at the location *(i,j)* in a *MxN* image. Subsequent to application of the FDR, the features were ranked according to their

class separability. The FDR scores are summarised in Table 1. The highest FDR scores were recorded for the B-blue feature values from the fillet RGB image, G-green feature values from the RGB image, and L-lightness feature values from the CIELab image. B, G, and L proved to be the most informative and most important features, resulting in the following feature R^3 space for a) classification between normal and defective fillets – $FColour = \{G_i,\ B_i, L_i\}$; and b) pixel-wise classification for blood spot segmentation $P_{(ij)} = (G_{ij}, B_{ij}, L_{ij})$. Based on the FDR score, the B, G and L features for discriminating between normal and defective fillets, and the B, G and L pixel level values for the pixel-wise classification of blood spot segmentation and localisation exhibited high between-class variance values and low values for within-class variance.

The fact that G-green values proved to be a good feature for blood detection is in good agreement with the absorption peak of blood at 577 nm [9], which is within the absolute wavelength proximity of green colour (Figure 2, Table 1) which is predominant in the wavelength range 495-570 nm [40]. As regards the B-feature taken from the RGB channel (Figure 2, Table 1), it has previously been shown that in general, myoglobin exhibits an absorption peak at 409 nm [41]. L –lightness (Figure 2) scored high on the FDR criterion for separability between normal and defective fillets (Table 1). Selection of the subset consisting only of 3 features (B, G and L, or "BGL") was considered optimal given the dataset available to this study. The *T/l* ratio between dataset sample size and the number of features used for the training of classification algorithms was greater than 20, which concurs with the recommendations reported in [27], and is in line with our aim to design a classification model with good generalisation properties.

3.2. Classification between normal and defective fillets

3.2.1 Support Vector Machines: The optimal BGL feature set combination for fillets in the training set enabled a linear separability to emerge between the classes in the 3D space as

defined by these features. The procedure to perform a grid search and validation of the test set was averaged for 100 randomised 80/20 splits (Table 2). This provided a measure of stability to the results which would not have been possible if only a single random split had been reported. The accuracy of the validation set was 99.5% (Table 2). Table 2 also shows the results from the exhaustive grid searches and selection of the optimal hyperparameters (C, γ pairs) resulting from the stratified 10-fold cross-validation, as well as for the wide grid search over a wide range of hyperparameters. We performed this operation to ensure that an exhaustive stratified 10-fold cross-validation was both sufficiently wide and fine-grained to capture all possible pairs of hyperparameters. Table 2 shows that the accuracy of the validation sets was over 99%, implying that from 100 random splits, only one fillet at one particular split was misclassified.

Table 2 Algorithm for classification task 1 and task 2. The first row shows the results for a narrow but fine-grained grid search (task 1). The second row shows the results of a wider and coarser grid search for task 1, with the third row displaying the results from task 2.

Classifier	Task	Kernel	Feature Set	Validation	Accuracy	Chosen C, γ	Training time (s)
C-SVM	1	RBF	G_m, B_m, L_m	80/20 x 100	99.5%	Ex: 2^{11}, 2^{-1}	4 secs.
C-SVM	1	RBF	G_m, B_m, L_m	80/20 x 100	99.46%	Ex: 2^{10}, 2^{-5}	4 secs.
C-SVM	2	RBF	G_{ij}, B_{ij}, L_{ij}	80/20	99.9%	2^{9}, 2^{-1}	48 mins.

3.2.2 CNN-based classification of normal and defective fillets

Figure 4 displays the AlexNet responses for a randomly chosen fillet image from the test data set. The AlexNet prediction exhibited a very high level of confidence (100%) in identifying this fillet as defective (blood spots). The original RGB input image at 256 x 256 resolution was converted to 227 x 227 x 3 as required by AlexNet [29]. The first convolutional layer (conv 1) filters the image using 96 kernels (layer depth 96) dimensioned 11 x 11 x 3, with a 4 pixel stride. Figure 4 also shows the respective learned filters applied to the fillet image. Each of the 96

filters is shared by the 55*55 neurons in one depth slice. The result of the first convolution layer is an activation map dimensioned 55 x 55. Each of the activation maps corresponding to the different filters are stacked to form an output map dimensioned 96x55x55 (conv1, Figure 4). Here it is shown how the blood spots are highlighted by some of the learned filters. We choose to visualize activation maps from selected layers and highlight in particular the activation maps for conv1, norm1, pool1, conv5, pool5 and 3 fully connected layers. This helps to give an understanding of how image data propagate in the layers of AlexNet. Activation maps from other layers not shown in Figure 4 are subsampling of activation maps from previous layers based on weights and trained parameters. The output of conv1, after normalisation and pooling, goes as input to the second convolutional layer conv2 and so on to generate the final prediction for the fully connected output layer fc8, whose binary activation map is shown in Figure 4. The model generated by fine-training the AlexNet was tested on 1300 unknown images, previously not included in the augmented training set. The model resulted in confident predictions of the class categories of the fillets. Table 3 provides a summary of the accuracy of the classification process, demonstrating a validation accuracy for 5108 images of 100%, the same as that for the 1300 test images. The prediction confidence for all 1300 test data set images was 100% for every single normal or defective fillet.

Table 3. Results for the classification model generated using AlexNet on a DIGITS 2.2 platform trained with n=15327 samples.

Classifier	**Task**	**Network**	**Validation (n=5108)**	**Accuracy (n=1300)**
Deep Learning	1	AlexNet	100%	100%

The data augmentation performed on our original dataset (Figure 3, see section 2.4.1) provides an effective means of compensating for the lack of large datasets in applications where the

476 acquisition of such data sets is a challenging and expensive process. This was our primary
477 motivation for employing data augmentation. Moreover, data augmentation is one of the easiest
478 and most common methods used to reduce overfitting [29]. It is well known that large size,
479 multi-parameter, networks can be prone to overfitting in situations involving limited sample
480 numbers in the training dataset [42]. Of particular interest in connection with our data
481 augmentation procedure was the process of filling in the black background of the fillet image
482 (Figure 2) with the respective RGB image mean (Figure 3). This study was not aiming to
483 investigate shape feature representations as such, but to reduce the sensitivity of the AlexNet to
484 fillet shape and increase its sensitivity to colour variations between the classes. As is
485 demonstrated in Figure 4, the AlexNet focus was shifted towards capturing intricate feature
486 representations based on the colour properties of the images that maximise the separability of
487 blood spots from muscle tissue. For this reason, many of the activation maps shown in Figure
488 4 contain less information than the learned filters can encode. It is seen how maps focusing on
489 the outer edges and shape of the fillet and exhibit very weak responses.

490 This indicates that the AlexNet possesses much greater descriptive power than is necessary
491 for the current application. However, this property can be beneficial for applications where
492 shape features are relevant for classification. There is a concern that the max-pooling of layers
493 may result in a loss of accuracy in spatial information [42] regarding the loss of valuable
494 localisation information especially in terms of detecting the precise spatial relationships
495 between the parts of a given object in connection with pose, orientation and scale in particular
496 [36]. For our application, we were not interested in pose, orientation or scale of the object and
497 this aspect was not relevant. Based on the results shown in Table 3, and as suggested in [17], it
498 seems that the good, intricate, features used to discriminate normal from blood spot fillets have
499 been automatically learned during training of AlexNet, resulting in good prediction accuracy
500 for the test data set in general terms.

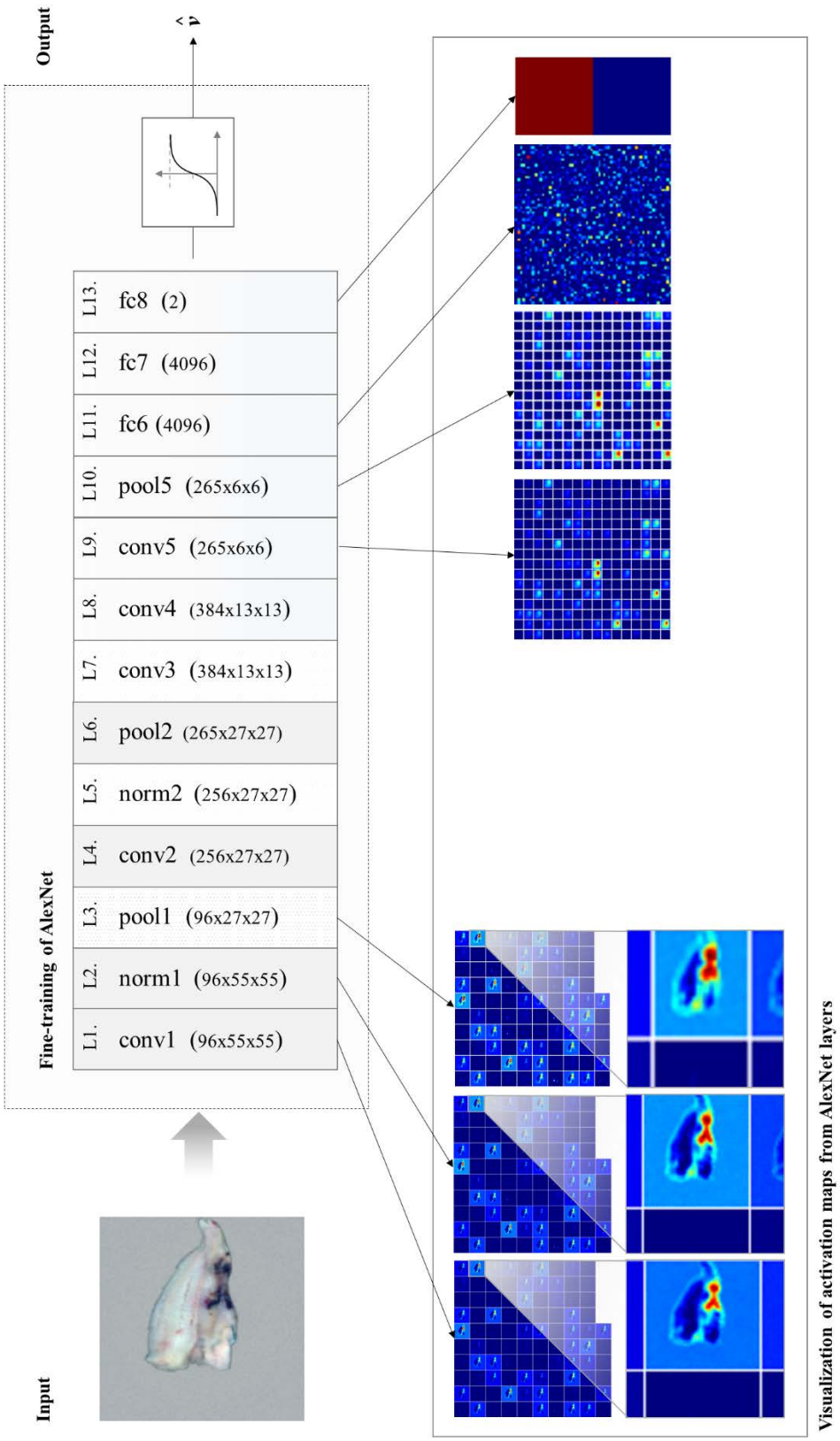


Figure 4: Visualization of activation maps in different layers of AlexNet showing theinput RGB image, activation maps for conv1, norm1, pool1, conv5, pool5, fully connected layer fc6 and fc8, as well as the close-up images of the fillet following convolution, contrast normalisation and pooling.

501

Results in Table 3 are comparable with results obtained from Kagaya et al. [18] in an application of recognizing food images. They used a 5-layer convolutional neural network for recognizing food images and they observed that the network achieved significantly better performance accuracy (93.8 %) than the baseline method (89.7%). Grinblat et al. [43] proposed an approach using CNN for the problem of plant identification from leaf vein patterns. The overall accuracy of implemented CNN models achieved significantly better accuracy on the test set compared to machine learning algorithms such as Support Vector machines or Penalized Discriminant Analysis, which is line with classification accuracy achieved in our CNN model outperforming the SVM model for the classification between normal and fillets with blood.

3.3. Pixel-wise classification for blood spot segmentation

This study adopted a similar approach for the pixel-wise classification for blood spot segmentation and localisation in the fillet image. As described in section 2.3.4, we used the pixel level colour values $P_{(ij)} = \left(G_{ij}, B_{ij}, L_{ij}\right)$ as features for the SVM classification model (Figure 1, Table 1). Individual pixels were randomly sampled, extracted and labelled according to information obtained from the ground truth images for known blood spots and muscle tissue regions. They were then classified. For those fillets exhibiting blood spots, a total of 130,000 pixels were labelled as belonging to blood spots. To reduce computational requirements, a fixed percentage of pixels was sampled from each class as training data. This resulted in the random sampling and extraction for training and validation of 20,000 pixels from the blood spot regions. Similarly, 20,000 pixels from normal muscle tissue regions were randomly extracted in order to obtain a balanced data set. It is known that random sampling is a simple method that minimises the effects of spatial pixel correlation [10], and for this reason the method is recommended for similar applications [44]. Figure 5 shows the resulting pixel-wise classification for blood spot segmentation from two randomly chosen images. Although the overall prediction accuracies for the validation set are consistently greater than 99% (Table 2,

third row), it should be noted that the training set samples are a subset of the full set of images. This means that an accuracy of 99.9% is a measure of the accuracy in detecting blood spots compared to the ground truth labelling. Labelled pixel samples defined as ground truth data are taken from regions classified as either “definitely blood” or “definitely normal muscle tissue”. This ‘conservative’ approach to the selection of ground truth data seemed unavoidable in situations where the blood spots gradually merge into muscle tissue.

Table 4. Blood segmentation performance evaluation for the proposed pixel-wise SVM classification algorithm using whole test fillet images.

Images	Accuracy	ER	TPR	TNR	time (ms)
All test images	95.41 %	4.59 %	99.48 %	95.38 %	1498
Image 13	98.98%	1.02%	100.00%	98.98%	1391
Image 21	96.36%	3.64%	100.00%	96.33%	1442

Table 4 presents the performance evaluation metrics (ER, TPR, TNR, CPU time) for blood segmentation from whole fillet images based on a pixel-wise SVM classification model applied to test images, and demonstrates that the proposed segmentation algorithm resulted in low ER, and high TPR and TPN values. The high TPR values (at or close to 100%), show that the algorithm classifies as “blood” all the pixels that were manually classified as “blood” from the ground truth data.

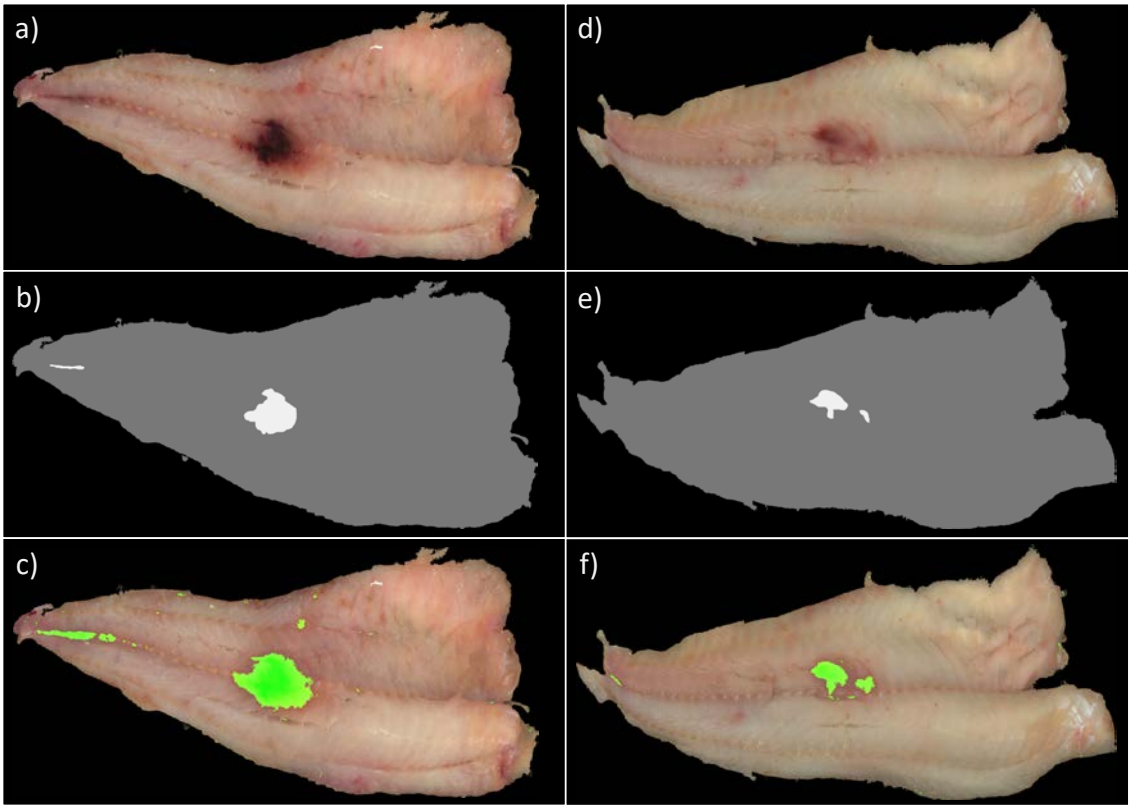


Figure 5. Resulting pixel-wise classification and segmentation of blood spots with the approach presented in this study (c, f). Blood segmentation for two fillet image examples (a, d). Comparison with ground truth data (b, e).

Our pixel-wise blood spot segmentation results and segmentation speed are comparable to the results reported in Mizushima et al. [45] regarding use of SVM for segmentation of apples. They achieved a 1.5 s segmentation average time for an apple on a platform consisting of iCore7, 3.4 GHz CPU, and 16 Gb RAM.

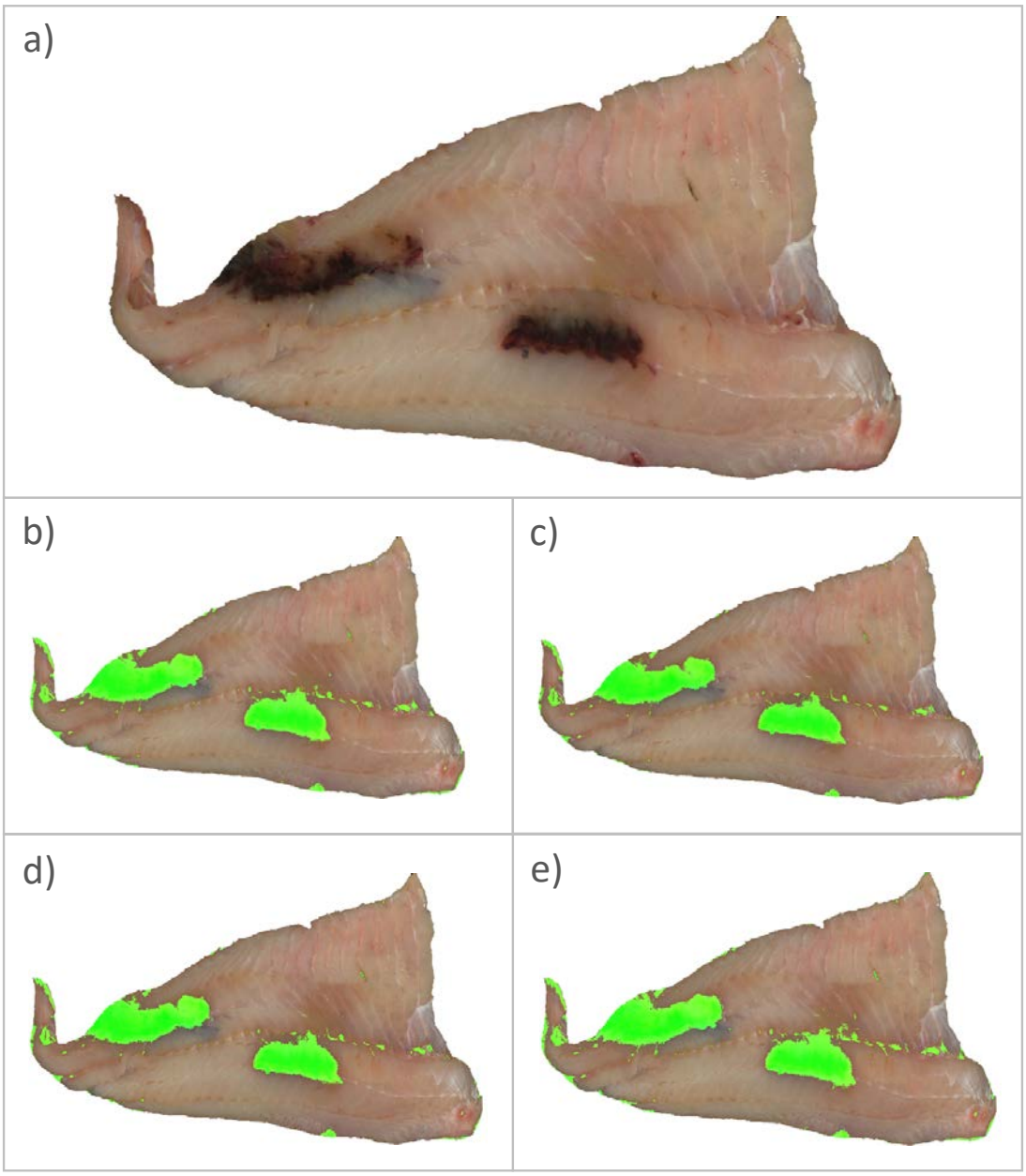
The false positives (mean FPR=4%) resulting from this pixel-wise classification come from two sources. Some are the result of conservative ground truth labelling and the means by which the lasso tool is applied along the blood spot boundary, while others are caused by the high spectral similarity between blood and other closely adjacent discolorations that appear in normal muscle tissue but which are not blood. This concurs with the findings in [44], according to which the presence of noise, combined with spectral similarities, represents the main cause of false classifications that occur when performing pixel-wise classifications based on spectral data. Both the labelled blood and normal muscle regions seem to share very similar colour

attributes as the blood spots gradually merge into normal muscle tissue. Given that our system and method are designed to be cost-efficient (based on RGB-D digital images acquired with a relatively low cost camera and computer vision set-up), the classification accuracy demonstrated in Table 4 is considered to be highly satisfactory for industrial purposes. Table 4 also shows that the proposed algorithm is time-efficient and highly relevant for rapid online industry applications, since the total mean run time required to segment a single fillet image once the algorithm is trained is 1.5 seconds. Table 4 also shows performance evaluation metrics for test image 13 (Figure 5, d and f) and image 21 (Figure 5, a and c). As suggested in [46], it seems that it is advantageous in the case of pixel-wise classifications for blood spot segmentation to operate with a low dimensional feature space using only 3 features; ($P_{(ij)} = (G_{ij}, B_{ij}, L_{ij})$. Keeping the number of features low also results in acceptable computational times (see Tables 2 and 3). Vempati et al [47] demonstrate that computational times using an RBF kernel increase linearly with data dimensionality, and non-linearly with the number of training samples.

3.4. The effect of SVM hyperparameters on pixel-wise classification and optimisation of classification performance

It is important here to highlight the effect of the choice of the hyperparameters (C and γ) on classification performance [48]. Figures 6, 7 and 8 illustrate the conceptual effect of changing hyperparameter values during the segmentation of blood spots from normal fillet muscle tissue. The classification model was trained using values of C and γ from opposite sides of the grid search. While both models yield the same overall classification accuracy in the grid search, the model trained with the lowest C value identifies a blood spot as extending over a larger region. This concurs with the observation that low C parameter values enable locations close to the boundary to be ignored, thus expanding the ‘margin’ region [49]. Figure 6c shows the pixel-wise classification for the same fillet as in Figure 6a using large values of C. In this case fewer

583 pixels are classified as ‘blood’ compared with the result in Figure 6b. Figure 7 provides a
584 visualisation of an isosurface computed for the RBF kernel decision values used in the applied
585 C-SVM algorithm. Figure 7a demonstrates that although a linear kernel is used, it is found that
586 a linearly separable data set is well suited to a linear separating hyperplane. Figure 7b
587 demonstrates that excessively high C values affect the decision isosurface by attempting to
588 classify all the data into the correct class. A major penalty must be paid for the use of high C
589 values in that the cost of misclassification is high, data points close to the hyperplane affect its
590 orientation, and the optimal separating hyperplane adopts a complex shape [48]. The C
591 parameter controls the loss attributed to samples that exceed the hyperplane margin, and can
592 therefore be used to fine-tune the decision boundary as blood spots merge into normal muscle
593 tissue. The kernel parameter γ also has a significant effect on the optimal separating
594 hyperplane/hypersurface. This parameter controls the width of the Gaussian kernel (σ) because
595 its relation with σ is given by $\gamma=1/(2\sigma^2)$. Figure 8 shows the effect of varying γ while keeping
596 the C ($C=2^1$) regularisation parameter constant during classification task 1.
597 At low γ parameter values (Figures 8d, 8e and 8f) the hyperplane is almost linear and exhibits
598 a low degree of curvature. As γ increases the curvature of the hyperplane changes (Figures 8a,
599 8b and 8c). Figure 8c shows that the decision hyperplane/surface has the largest curvature when
600 $\gamma=2^5$, because the decision surface is forced to curve in order to avoid misclassifications, thus
601 introducing the risk of overfitting as is also described in [48]. But what is the effect of changes
602 in the γ parameter for blood segmentation in the pixel-wise classification?



603

604 **Figure 6.** The effect of changing the hyperparameters C and γ on the classification accuracy of blood

605 segmentation of an example fillet (a); varying the C parameter (b, c), keeping C constant ($C=2^1$) but changing γ

606 between $\gamma=2^5$(d), and $\gamma=2^{-5}$ (e). The value of $\gamma = 2^5$ results in an accuracy of 96%, 100% true positive rate,

607 and 95.5% true negative rate.

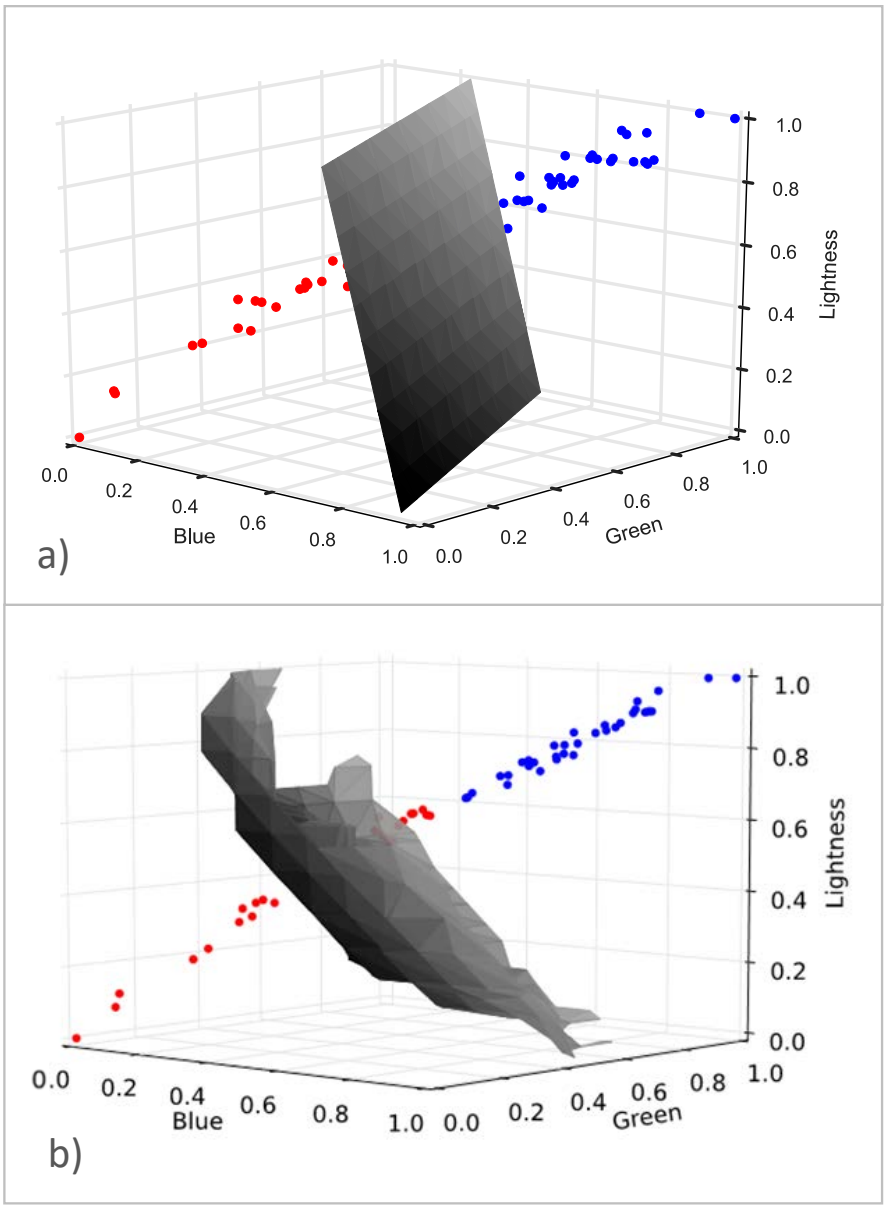


Figure 7. The decision hyperplane visualisation for the SVM algorithm using an RBF kernel. a) Although the RBF kernel is used for a linearly separable dataset, a linear optimal separating hyperplane is found to be suitable ($C=2^{15}, \gamma = 2^{-5}$); b) A separating isosurface with a change of orientation for $C=2^{50}, \gamma = 2^{-5}$ as an example of a poor classifier.

Figures 6d and 6e shown pixel-wise classification results using two different values of γ, while keeping the regularisation parameter C constant. The example in Figure 6d uses a kernel where $\gamma = 2^5$, while that in Figure 6e uses $\gamma = 2^{-5}$. If we define the kernel width as "Full Width at Tenth Maximum" (FWTM), the calculation $W = 2\sqrt{2\ln(10)}\sigma$=4.291σ [50] can be performed. For the example shown in Figure 6d, the kernel width is $W_1 = 17.164$,

while for that in Figure 6e, the width is $W_2 = 0.5359$.

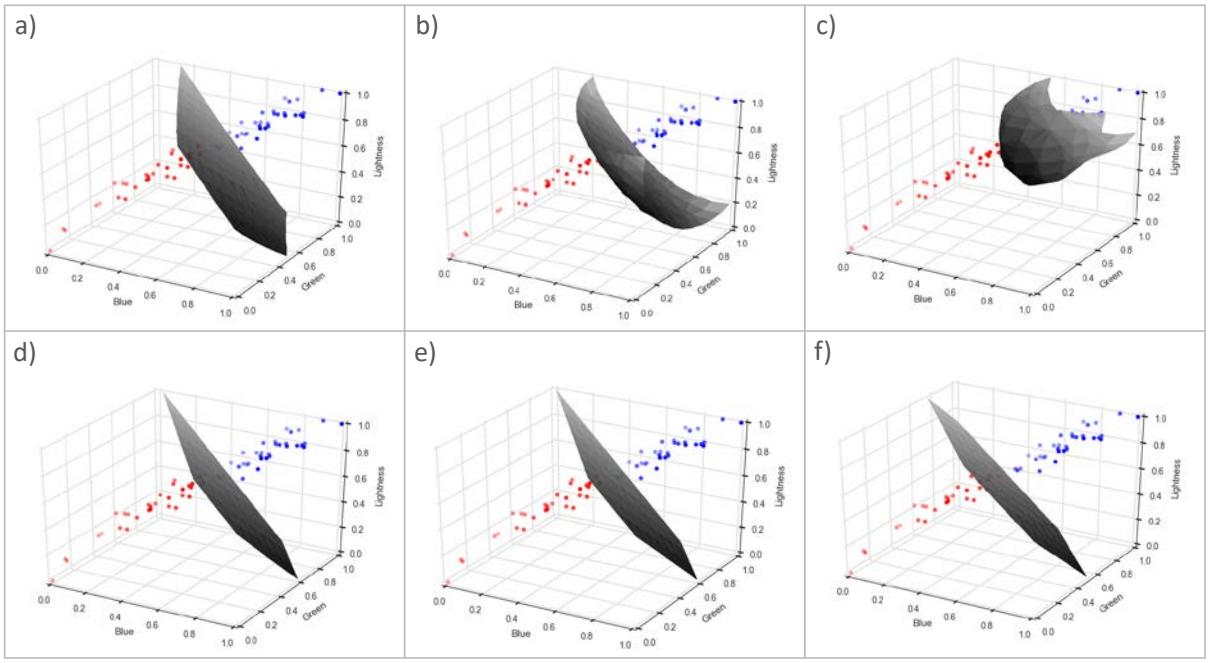


Figure 8. Illustration of the effect of changing the $\boldsymbol{\gamma}$ parameter in the decision hyperplane/surface for the RBF kernel. a) $\boldsymbol{\gamma = 2^{1}}$; b) $\boldsymbol{\gamma = 2^{2}}$; $\boldsymbol{c}$) $\boldsymbol{\gamma = 2^{5}}$; d) $\boldsymbol{\gamma = 2^{-1}}$; e) $\boldsymbol{\gamma = 2^{-2}}$; f) $\boldsymbol{\gamma = 2^{-5}}$. Figure 8 (c) shows that high values of $\boldsymbol{\gamma}$ change the curvature of the decision hyperplane/surface in an attempt to avoid pixel mis-classification.

For the image displayed in Figure 6d, the resulting classification exhibited an overall accuracy of 96%, 100% true positive rate, and 95.5% true negative rate (slightly higher – 1% – than the classification using $\gamma = 2^{-5}$). It is clear that, although lower γ parameter values result in slightly lower levels of accuracy (94,5% for $\gamma = 2^{-5}$), the classification model in these cases is considered to exhibit better generalisation properties in terms of the classification of unknown samples, whereas the use of high γ values may introduce the risk of overfitting and poor generalisation.

3.5. Blood spot segmentation and localisation in 3D space and calculation of gripper vectors for robotic processing

Figure 9 demonstrates how the results obtained can be used for the accurate localisation of blood spots in 3D space. Perfectly per-pixel aligned RGB and 3D images of the fillet are shown with the blood spots segmented from normal muscle tissue, oriented in an OXYZ frame. The

resulting RGB image from the pixel-wise classification is combined with the 3D image of the fillet, and a RGB-D map (in mm) is generated in order to achieve accurate localisation of blood spots in terms of OXYZ coordinates. A normal gripper vector is calculated for each blood spot with respective vector origin coordinates (in mm) relative to the origin of the OXYZ frame where P_1 (125.7, 123.7, 16) is used for gripper vector 1, and P_2 (284.7, 157.2, 26) for gripper vector 2. This information is sufficient for a robot manipulator to perform automated trimming of the blood spots. Once control of the 3D coordinates of the entire blood spot regions has been achieved, it is a straightforward task for any robotic gripper or cutting tool to remove blood spots because the gripper motion path can now follow the 3D profile of the fillet to the specific area marked as a blood spot given in OXYZ coordinates. This demonstrates how classification results can be directly converted into information that is relevant to automated robotic processing.

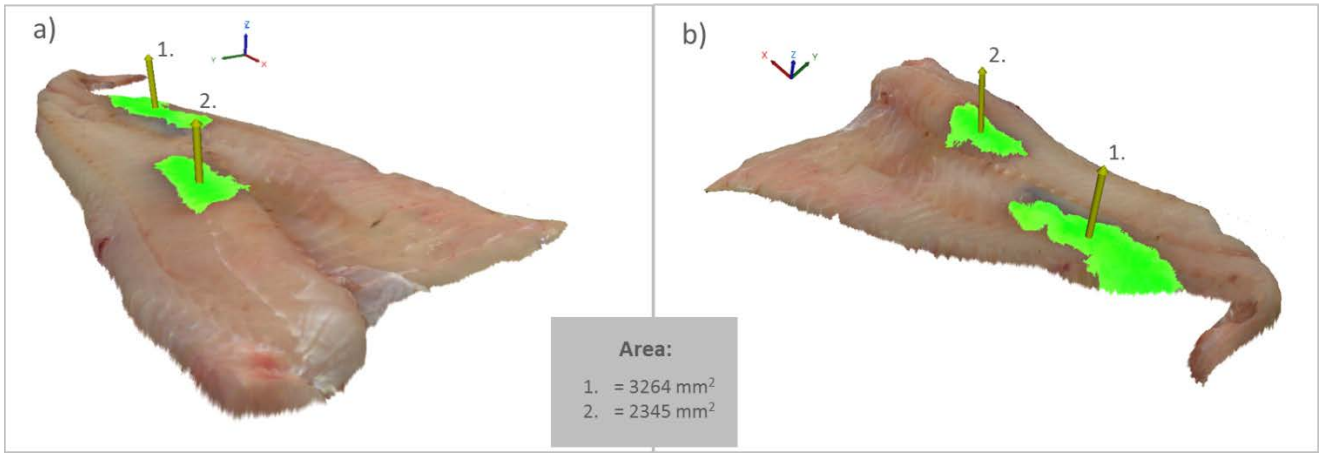


Figure 9. Visualisation of blood segmentation and blood spot localisation in 3D space (OXYZ coordinates). The visualisation of 3D gripper vectors normal to the plane is defined by the presence of blood spots and a specification of the area (in mm) covered by spots.

3.6 Future work and future research directions

As future work, several approaches might be considered in order to increase the robustness of methods used to identify defective fillets, and of pixel-wise classification approaches used for the segmentation of blood spots based on RGB-D images. Studies addressing texture

classification and colour image segmentation [10, 51] have shown that it is possible to include textural features to improve pixel-wise classification models. However, these methods are best suited to applications where the texture of the objects or regions in question exhibits high levels of contrast. They are less well suited to dealing with biological raw materials that exhibit low variations in texture since a typical cod fillet will exhibit very homogeneous muscle tissue patterns regardless of the presence of blood spots.

Based on the presented research results and observed trends from the bibliography we foresee an increased focus on the synergies of computer vision, deep learning and robotics also for food, biomarine and agricultural applications resulting in future research directions:

- Increased use of 3D information and combined RGB-D images in inspection, recognition, and robotic application tasks. 3D is invaluable information for detection, recognition and localization of objects in the scene and localization, in 3D space, of the defects in the object itself. For example, we use the 3D information to localize in 3D space the blood spots and to calculate the relevant gripper vectors. Similarly, the 3D information can be used for localization of gaping in the cod fillets. The advent of RGB-D cameras such as Kinect v2, Intel RealSense SR300 and specialized hardware from manufacturers such as NVIDIA and Intel will open up for new research developments in using 3D information for development of novel and robust machine learning models.
- Transfer learning – increased use of the existing pretrained CNN architectures on large datasets, such as AlexNet, VGG16, VGG19, and fine-tuning of these networks for various inspection, recognition, robotic application tasks. This is because in food, biomarine and agricultural domains the datasets are limited and it is challenging to acquire large enough datasets to train the network architectures from the scratch. Another modality is to use the CNN as a feature extractor by removing the last fully

connected layer. In this way the CNN would generate automatically learned features alleviating the need for hand-engineered features. In our case, we used a pre-trained AlexNet on a large dataset, and fine-tuned the network with our particular dataset of cod fillets consisting of normal fillets and those with blood spots.

- In applications where temporal aspect is important such as action recognition of livestock from video, there will be an increased use of CNN and Recurrent Neural Networks - RNN in combination with LSTM-Long Short Term Units and GRU-Gate recurrent units. LSTM or GRU units add a memory component to the network which is important to capture motion features for action recognition. The application domains of such network architectures can be action recognition of livestock in order to estimate the welfare status or to optimize operations such as feeding. Concretely such architectures can be used for study of fish behaviour and action recognition from underwater video.
- Robot learning – Robot 3D manipulation, handling and processing of complex agricultural products is still challenging due to, among others, inability to teach robots to dynamically manipulate the raw material. Increased focus on use of deep learning architectures to process visual information and learning to grasp, manipulate, and process objects is necessary to improve the performance of robots. One strategy for learning is the Learning from Demonstration where humans guide the robot, either physically or by teleoperation, to teach a skill. This could be an example strategy to use for learning the robot to trim the blood spot from the fillet. However, the robot through Learning from demonstration can only be as good as the teacher and in some other applications other learning strategies should be considered. Deep reinforcement learning is, for example, one learning strategy we are going to see more in robotic agricultural applications due to ability to endow robots with manipulation and

behavioural skills without much human intervention and with possibility to improve the learned behaviour and skill over time.

- Synthetic generation of image datasets for training of classification and prediction models – in cases and applications where acquiring large datasets is challenging. It is known that acquisition of large datasets in food and agricultural industries is limited and very often it is unpractical and very expensive to acquire such datasets.

4. Conclusions

In this study, we present approaches for classification between normal and defective fish fillets based on SVM classifier and GPU-accelerated convolutional neural networks, together with a robust SVM- pixel-wise classification approach for blood spot segmentation based on RGB and 3D images. The best hand-engineered features for optimisation of the discrimination of normal muscle tissue from blood spots were found to be G-green, B-blue, and L-lightness image features in RGB and CIELab colour space. A summary of main conclusions regarding the main research objectives is as follows:

a) Development of robust, colour-based pixel-wise classification for blood spot segmentation in cod fillets: The pixel-wise classification model, employing a SVM algorithm with a Gaussian RBF kernel using pixel level features resulted in an overall classification accuracy of 99%, 99.5% for sensitivity, and 95.4 for specificity. The SVM based pixel-wise model demonstrates that results can be used for accurate segmentation and localization in 3D space and calculation of respective gripper vectors for robotic processing of such defects and similar biological raw muscle tissue where defective and normal regions exhibit high levels of spectral similarity.

b) Classification of normal and defective fillets: Both SVM and CNN-based models showed good classification accuracies for the test sets with CNN-model slightly outperforming the SVM-model (100% vs 99%). The results from these models show

that this approach to the classification may have applications beyond the specific scope of this study.

c) Per-pixel aligned RGB-D images: the approach results in perfectly per-pixel aligned RGB and D images of high resolution acquired in real-time during scanning of the fillet while it is transported on conveyer belt.

d) Conceptual effect of the SVM hyperparameters: a visualization of the change of SVM hyperparameters C and γ gives a better understanding on these two parameters and their effect on the classification accuracy. This is important in order to prevent model overfitting.

e) CNN capabilities for classification of food objects and correct data augmentation: The deep learning approach implemented by fine-tuning of the pretrained AlexNet resulted in 100% classification accuracy between normal and defective fillets. The data augmentation approach to desensitize the CNN for shape and focus only on colour features resulted in high classification accuracy between normal and defective food products. As a result of desensitization process, many activation maps of the AlexNet contained less information than the learned filters can encode. Despite this, the results show that AlexNet possesses much greater descriptive power than it was necessary for our application in classification of cod fillets.

The proposed approaches, although demonstrated in laboratory scale, have also practical industrial relevance given segmentation of blood spots by the pixel-wise classification model is rapid, with possibility to time-optimize it, as it is currently able to process one fillet image in average 1.5 seconds. This opens up for potential real-time industrial use of the reported approaches. For transfer of these methods to industrial application, software optimization with regard to increased speed of operation is necessary, in addition to complying with hardware requirements deriving from operating in humid and cold production environment.

Acknowledgments

This study was funded by the Research Council of Norway as part of the research project QualiFish (Project number 233709; http://qualifish.no/). This is a multidisciplinary research project funded via the BIONÆR program (www.forskningsradet.no/bionaer). Aleksander Eilertsen (MSc.) is gratefully acknowledged for his work on the figure illustrations.

References

[1] U. Erikson, E. Misimi, B. Fismen, Bleeding of anaesthetized and exhausted Atlantic salmon: body cavity inspection and residual blood in pre-rigor and smoked fillets as determined by various analytical methods, Aquaculture Research. 41(4) (2010) 496-510.

[2] H.Y. Yang, X.J. Zhang, X. T. Wang, LS-SVM-based image segmentation using pixel color-texture descriptors, Pattern Anal Appl. 17 (2014) 341-359.

[3] E. Misimi, J.R. Mathiassen, U. Erikson, Computer Vision Sorting of Atlantic Salmon (Salmo salar) Fillets According to Their Color Level, J Food Sci. 72(1)(2007) S030-S035.

[4] P. Jackman, D.-W. Sun, C.-J. Du, P. Allen, Prediction of beef eating qualities from color, marbling and wavelet surface texture features using homogenous carcass treatment, Pattern Recogn. 42(5) (2009) 751-763.

[5] M.O. Balaban, G.F. Ünal Şengör, M.G. Soriano, E.G. Ruiz, Quantification of gaping, bruising, and blood spots in salmon fillets using image analysis, J Food Sci. 76(3) (2011) E291-7.

[6] E. González-Rufino, P. Carrión, E. Cernadas, M. Fernández-Delgado, R. Domínguez-Petit, Exhaustive comparison of colour texture features and classification methods to discriminate cells categories in histological images of fish ovary, Pattern Recogn. 49(9) (2013) 2391-2407.

[7] E. Cernadas, P. Carrion, P.G. Rodriguez, E. Muriel, T. Antequera, Analyzing mangentic resonance images of Iberian pork loin to predict its sensorial characteristics. Computer Vision and Image Understanding 98 (2005) 345-361.

[8] N. Martinel, C. Piciarelli, C. Micheloni, A supervised extreme learning committee for food recognition, Computer Vision and Image Understanding. 148(2016) 67-86.

[9] K. Mertens, B. Kemps, C. Perianu, J. Baerdemaeker, E. Decuypere, B. Ketelaere, M. Bain,. Advances in egg defect detection, quality assessment and automated sorting and grading, in: Y. Nys, M. Bain, F. Immerseel, (Eds.), Improving the safety and quality of eggs and egg products. Volume 1: Egg chemistry, production and consumption, 2011, pp. 209-241.

785 [10] X.-Y. Wang, T. Wang, J. Bu, Color image segmentation using pixel wise support vector machine
786 classification, Pattern Recogn. 44 (2011) 777-787.
787 [11] R. Unnikrishan, C. Pantofaru, M. Hebert, Toward Objective Evaluation of Image Segmentation Algorithms,
788 IEEE Trans Pattern Anal Mach Intel. 29(6) (2007) 929-944.
789 [12] V. Vapnik, The nature of statistical learning theory. New York: Springer-Verlag, 1995.
790 [13] S. Deng, Y. Xu, L. Li, X. Li, Y. He, A feature-selection algorithm based on Support Vector Machine-Multiclass
791 for hyperspectral visible spectral analysis, J Food Eng. 119(1) (2013) 159-166.
792 [14] J. Rousu, L. Flander, M. Suutarinen, K. Autio, P. Kontkanen, A. Rantanen, Novel computational tools in
793 bakery process data analysis: A comparative study. Journal of Food Engineering, 57(1) (2003) 45–56.
794 [15] X. Sun, K. J. Chen, K.R. Maddock-Carlin, V. L. Anderson, A.N. Lepper, C.A. Schwartz, W.L. Keller, B.R.
795 Ilse, J.D. Magolski, E.P. Berg, Predicting beef tenderness using color and multispectral image texture features,
796 Meat Science. 92(4) (2012) 386-393.
797 [16] C.-J. Du, D.-W. Sun., Pizza sauce spread classification using colour vision and support vector machines.
798 Journal of food engineering 66 (2005) 137-145.
799 [17] Y. LeCun, Y. Bengio, G.E. Hinton, Deep Learning, Nature. 521 (2015) 436-444.
800 [18] H. Kagaya, K. Aizawa, M, Ogawa, Food detection and recognition using convolutional neural network. MM
801 '14 Proceedings of the 22nd ACM international conference on Multimedia, pp. 1085-1088, Orlando, Florida,
802 USA.
803 [19] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016, pp. 233-235.
804 [20] E. Misimi, E.R. Øye, A. Eilertsen, J. R. Mathiassen, T. Gjerstad, O. Åsebo, J. Buljo, Ø. Skotheim. Gribbot -
805 Robotic 3D vision-guided harvesting of chicken fillets, Computers and electronics in agriculture 121 (2016)
806 84-100.
807 [21] M.O. Balaban, E. Misimi, Z. Alcicek, Quality Evaluation of Seafoods, in: D.-W. Sun (Eds.), Computer Vision
808 Technology for Food Quality Evaluation, Elsevier, 2016. pp. 243-271.
809 [22] F. Mendoza, P. Dejmek, J.M. Aguilera, Calibrated color measurements of agricultural foods using image
810 analysis, Postharvest Biol. Technol. 41 (2006) 285–295.
811 [23] U. Erikson, E. Misimi, Atlantic Salmon Skin and Fillet Color Changes Effected by Perimortem Handling
812 Stress, Rigor Mortis, and Ice Storage, J. Food Sci.. 73 (2008) C50-C59.
813 [24] H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-
814 relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intel. 27(8) (2005) 1226-1238.

[25] S. Theodoridis, K. Koutroumbas, Pattern Recognition. 4th ed., Academic Press, Burlington, MA, 2009.

[26] E. Misimi, U. Erikson, A. Skavhaug, Quality Grading of Atlantic Salmon by Computer Vision, J. Food Sci. 73(5)(2008) E211-E217.

[27] S. Finette, A. Bleier, W. Swindel, Breast tissue classification using diagnostic ultrasound and pattern recognition techniques: I. Ultrasonic Imaging, 5 (1983) 55–70.

[28] Y. LeCun, K. Kavukcuoglu, C. Farabet, 2010. Convolutional Networks and Applications in Vision. In Proceedings of ISCAS - The IEEE International Symposium on Circuits and Systems, 2010, pp. 253-256.

[29] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems 25 (NIPS 2012).

[30] X. Glorot, A. Bordes, Y. Bengio, Deep sparse rectifier neural networks. J Mach Learn Res. (2011) 315-323.

[31] G. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, Improving neural networks by preventing co-adaptation of feature detectors. http://arxiv.org/abs/1207.0580, 2012.

[32] C.-C. Chang, C. Lin, LIBSVM: A Library for Support Vector Machines, https://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[33] C.-W. Hsu, C.-C. Chang, C.-J. Lin, A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University. 2003. http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.

[34] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, 2nd Edition, John Wiley&Sons, New York, 2001.

[35] C.-W. Hsu, C.-J. Lin, A Simple Decomposition Method for Support Vector Machines, Mach Learn, 46 (2002) 291-314.

[36] C. Huang, ACO-based hybrid classification system with feature subset selection and model parameters optimization. Neurocomputing, 73(1-4) (2009) 438-448.

[37] S.L. Salzberg, On comparing classifiers: Pitfalls to avoid and a recommended approach, Data Mining and Knowledge Discovery. 1 (1997) 317-327.

[38] C.-W. Hsu, C.-J. Lin, A comparison of methods for multiclass support vector machines, IEEE Trans. Neural Networks. 13(2) (2002) 415–425.

[39] J. Bergstra, Y. Bengio, Random Search for Hyper-Parameter Optimization, J Mach Learn Res. 13(2012) 281-305.

[40] T.J. Bruno, P.D.N. Svoronos, CRC Handbook of Fundamental Spectroscopic Correlation Charts, CRC Press, 2005, p. 2.

[41] P.A. Sykes, H.-C. Shiue, J. R. Walker, R.C. Bateman Jr., Determination of Myoglobin Stability by Visible Spectroscopy, J Chem Educ. 76(9) (1999) 1283-1284.

[42] C. Szegedy, W. Lu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich. Going deeper with convolutions, in: Proceedings of Computer Vision and Pattern Recognition CVPR-Conference, 2015.

[43] G.L., Grinblat, L. C. Uzal, M.G. Larese, P. M. Granitto, Deep learning for plant identification using vein morphological patterns. Computers and electronics in agriculture 127 (2016) 418-424.

[44] L. Paluchowski, E. Misimi, L. Grimsmo, L.L. Randeberg, Towards automated sorting of Atlantic cod roe, milt, and liver – Spectral characterization and classification using visible and near-infrared hyperspectral imaging, Food Control. 62(2016) 337-345.

[45] Mizushima, A., Lu, R. An image segmentation method for apple sorting and grading using support vector machine and Otsu's method. Computers and electronics in agriculture 94: 29-37.

[46] O. Chapelle, P. Haffner, V. Vapnik, SVMs for Histogram-Based Image Classification, IEEE Trans Neural Netw. 10(5) (1999) 1055-1064.

[47] S. Vempati, A. Vedaldi, A. Zisserman, C. V. Jawahar, Generalized RBF feature maps for Efficient Detection, in: F. Labrosse, R. Zwiggelaar, Y. Liu, B. Tiddeman (Eds.), Proceedings of the British Machine Vision Conference, BMVA Press, 2010, pp. 2.1-2.11.

[48] A. Ben-Hur, J. Weston, A User's Guide to Support Vector Machines, in: O. Carugo and F. Eisenhaber (Eds.), Biological Data Mining. Springer Protocols, 2009, pp. 223-239.

[49] S.S. Keerthi, C.-J. Lin, Asymptotic behaviours for Support Vector Machines with Gaussian Kernel, Neural Computation. 15(7) (2003) 1667-1689.

[50] O. Demirkaya, M.H. Asyali, P.K. Sahoo, Image Processing with MATLAB: Applications in Medicine and Biology (MATLAB Examples). CRC Press, Boca Raton, FL, 2008, pp. 380-383.

[51] K.I. Kim, K. Jung, S.H. Park, H.J. Kim, Support Vector Machines for Texture Classification, IEEE Trans Pattern Anal Mach. Intell. 22(11) (2002) 1542-1550.