





# Discriminating healthy from tumor tissue in breast lumpectomy specimens using deep learning-based hyperspectral imaging

LYNN-JADE S. JONG,<sup>1,2,6</sup>  NAOMI DE KRUIF,<sup>3,6</sup> FREIJA GELDOF,<sup>1,2</sup> DINUSHA VELUPONNAR,<sup>1,2</sup> JOYCE SANDERS,<sup>4</sup> MARIE-JEANNE T. F. D. VRANCKEN PEETERS,<sup>1</sup> FREDERIEKE VAN DUIJNHOFEN,<sup>1</sup> HENRICUS J. C. M. STERENBORG,<sup>1,5</sup> BEHDAD DASHTBOZORG,<sup>1,3,\*</sup>  AND THEO J. M. RUERS<sup>1,2</sup>

<sup>1</sup>Department of Surgery, Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands

<sup>2</sup>Faculty of Science and Technology, University of Twente, Drienerlolaan 5, 7522 NB Enschede, The Netherlands

<sup>3</sup>Department of Biomedical Engineering, Eindhoven University of Technology, PO Box 513, 5600 MB Eindhoven, The Netherlands

<sup>4</sup>Department of Pathology, Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands

<sup>5</sup>Department of Biomedical Engineering and Physics, Amsterdam University Medical Center, Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands

<sup>6</sup>Equal contributors

\*[b.dasht.bozorg@nki.nl](mailto:b.dasht.bozorg@nki.nl)

**Abstract:** Achieving an adequate resection margin during breast-conserving surgery remains challenging due to the lack of intraoperative feedback. Here, we evaluated the use of hyperspectral imaging to discriminate healthy tissue from tumor tissue in lumpectomy specimens. We first used a dataset obtained on tissue slices to develop and evaluate three convolutional neural networks. Second, we fine-tuned the networks with lumpectomy data to predict the tissue percentages of the lumpectomy resection surface. A MCC of 0.92 was achieved on the tissue slices and an RMSE of 9% on the lumpectomy resection surface. This shows the potential of hyperspectral imaging to classify the resection margins of lumpectomy specimens.

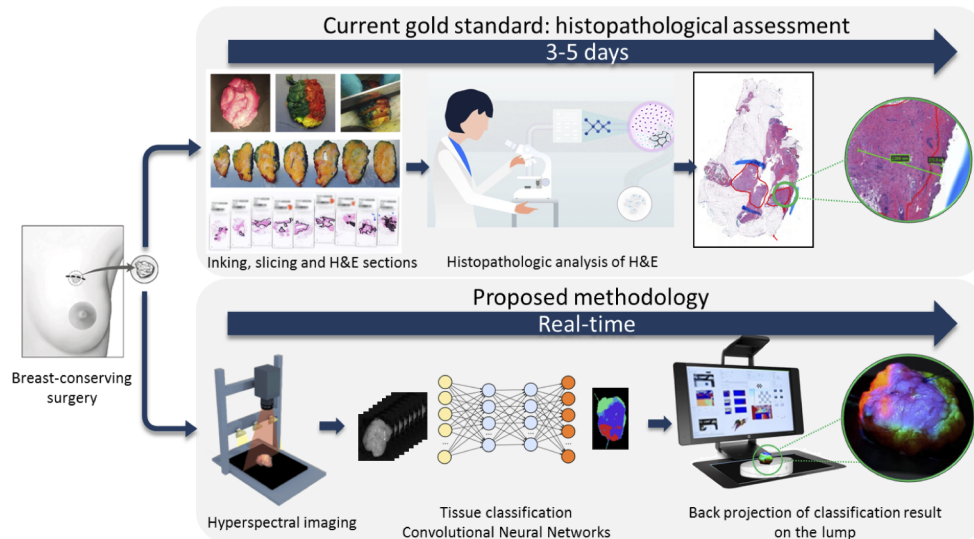
© 2022 Optica Publishing Group under the terms of the [Optica Open Access Publishing Agreement](#)

## 1. Introduction

Breast cancer is the most common cancer among women worldwide with nearly 2.3 million new cases diagnosed in 2020 [1]. The standard treatment for early-stage breast cancer is surgical resection of malignant tumor tissue by breast-conserving surgery (BCS) [2,3]. With a BCS the surgeon aims to remove the tumor entirely with a small margin of healthy tissue while preserving the breast as much as possible.

After surgery, the margins of the resected tissue are investigated to assess whether the tumor was completely removed. This is done by histopathologic analysis (Fig. 1), which is currently the gold standard for assessing surgical tissue. According to guidelines of the Society of Surgical Oncology (SSO) and the American Society for Radiation Oncology (ASTRO) a margin is defined as tumor-positive when there is invasive carcinoma (IC) on the resection surface (i.e. ink on tumor) or ductal carcinoma in situ (DCIS) within a 2 mm distance from the surface [4]. Patients with an incomplete tumor removal (i.e. a positive resection margin) often require a re-excision or boost radiotherapy to clear residual malignant tissue and to prevent cancer recurrence. According to the eusomaDB database, a re-excision was needed in 27% of the surgeries to achieve adequate

margins [3]. Such additional treatments increase the medical cost, negatively affect the cosmetic outcome and the patients' quality of life [5].



**Fig. 1.** The upper pipeline shows the conventional histopathologic analysis, which takes 3 to 5 days to process the resected breast tissue into hematoxylin and eosin (H&E) stained sections and to assess the resection margins. The lower pipeline demonstrates our envisaged method for real-time intraoperative feedback during breast-conserving surgery by means of hyperspectral imaging, deep neural networks and a projection mapping system.

Despite the dramatic improvement in preoperative imaging in healthcare made over the last decade, the surgeon still has to rely on visual and tactile feedback to distinguish malignant tumor tissue from healthy tissue during BCS. On top of that, histopathological assessment of the resection margins requires 3 to 5 days. Consequently, no feedback can be given to the surgeon during surgery and therefore there is a need for a margin assessment technique that can provide accurate intraoperative feedback about the entire resection margin in a limited amount of time. In this way, immediate action can be taken by the surgeon to still guarantee complete tumor removal.

The margin assessment techniques that are currently available in the clinic are frozen section analysis, imprint cytology and specimen radiography. However, these techniques either exhibit a low accuracy or are too time-consuming to examine the entire resection surface during surgery [6].

Due to these disadvantages, a variety of imaging and spectroscopy methods were proposed: ultrasound [7], radiofrequency spectroscopy [8], Raman spectroscopy [9], diffuse reflectance spectroscopy [10] and optical coherence tomography [11]. Studies showed that these techniques achieve a sensitivity from 70 to 100% and specificity from 67 to 93%. Despite the potential, these techniques have various practical drawbacks, such as a small field-of-view and an excessive time to analyze the entire resection surface.

In this work, we aim to develop an innovative method for tumor detection in the resection plain during surgery that overcomes all the limitations of the current technologies (Fig. 1). To this end, we investigated hyperspectral (HS) imaging as an intraoperative margin assessment technique on the removed tissue immediately after resection. HS imaging is a novel optical imaging technique that can image the spectral properties of a large surface in a short time without requiring any contact or the administration of contrast agents. By imaging the diffuse reflected light over a broad wavelength range, the intrinsic optical properties of the tissue's entire resection surface

can be measured and stored in a 3D hypercube, containing both the spectral and spatial data of the tissue. Since the optical properties depend on the tissue's composition and morphology, they are characteristic for each tissue type and can be therefore used to discriminate malignant tumor from healthy tissue [12–14]. HS imaging will be performed in the OR so that real-time results can be shown to the surgeon during surgery. This enables immediate surgical re-excision when indicated.

In previous studies at the Netherlands Cancer Institute - Antoni van Leeuwenhoek Hospital, an extensive dataset is created on optical properties of *breast tissues slices* using HS imaging [15,16]. In the paper of Kho *et al.* [16] two different classifications are used for the discrimination of tissue types. The first classification is only based on spectral information using a Fisher's linear discriminant analysis (LDA) classifier and the second one is a deep learning-based technique using both spectral and spatial information [16]. Both methods can distinguish the tissue types on the inked and gross-sectioned breast tissue slices with a sensitivity and specificity higher than 76%. However, using both spectral and spatial information resulted in a better performance.

With the high potential of HS imaging in discriminating tumor from healthy tissue shown in breast tissue slices, we are now aiming to image *unprocessed whole surgical resection specimens*, also called "lumpectomy specimens", instead of tissue slices. This is necessary to be able to assess the resection margins of the excised breast tissue in real-time. With the tissue slices we could only image the inside of the breast tissue after being sliced at the pathology department. However, with the lumpectomy specimens on the other hand, we would be able to image the original resection plain within the surgical workflow in the operation theatre. By analyzing the HS images with convolutional neural networks and incorporating both spectral and spatial information, we could thus provide real-time feedback on the margin assessment and tissue type classification of such lumpectomy specimens (Fig. 1).

Machine learning algorithms are proposed as a technique for HS imaging classification, for example, Support Vector Machine, logistic regression, and k-Nearest neighbors [17]. However, these methods require preprocessing of the raw data to extract hand-crafted features, thus expertise of the raw data is necessary. In contrast, deep learning models can automatically extract an effective feature representation of the raw input data, which is the main advantage of this machine learning subcategory. The models have a hierarchical structure, where the bottom layers can extract low-level features. Those low-level features are used as input for the top layers to create high-level features that are more abstract and discriminative [18]. Deep learning models that are proposed for analyzing HS images include stacked autoencoders, deep belief networks and convolutional neural networks [17].

In particular, convolutional neural networks (CNNs) have shown great classification performance on HS imaging [19–27]. Compared to stacked autoencoders or deep belief networks, the CNNs can extract features from images without resizing them into a 1D vector and thus both spectral and spatial information will be retained. Li *et al.* [23] proposed a 3D-CNN framework for HS imaging land cover classification, and compared this method with other deep learning methods, such as stacked autoencoder, deep belief network and 2D-CNN. Results show that their 3D-CNN model achieved a higher performance. Similar work on geoscience and remote sensing was performed by Chen *et al.* [19]. The authors applied a 3D-CNN-based network to extract spectral-spatial features and compared this network with a 1D-CNN and 2D-CNN, where respectively only spectral or spatial features were extracted. Results indicated that the network achieved a better performance using both spectral and spatial features. In addition, the authors investigated some methods to reduce overfitting, including L2 regularization and dropout. A dual-channel network for land cover classification was proposed by Yang *et al.* [26] that exploits the spectral and spatial information in two separate channels. The first channel used 1D convolution to extract spectral features from the pixel spectrum and the second channel used 2D convolution to extract spatial features from the average of the spectral bands. Both features

were concatenated and subsequently fed into the fully connected layers to extract spectral-spatial features for classification. Although the recently proposed CNNs have shown good performance for the classification of HS images, they were used for non-medical applications. When applying to medical HS data, one major challenge will be the limited number of labeled samples for training a similar classification algorithm. In particular, by using CNNs to make the ultimate step towards HS imaging analysis of lumpectomy specimens, we expect to encounter the following main challenges:

- For a "lumpectomy dataset", the samples are imaged immediately after resection, and due to logistical reasons as set by the histopathological protocol, only up to three locations can be marked with black histopathology ink, which remains visible on the H&E sections under the microscope. These selected three locations only cover a small fraction of the entire surface that is measured with HS imaging, and are in strong contrast to the breast tissue slices analysis where the whole HS image can be registered to the H&E section. So, due to this histopathological protocol which has to be followed to comply with clinical standards, it is difficult to correlate histopathology results with corresponding acquired HS images from the surface of lumpectomy specimens for each pixel. Consequently, we can only obtain the ground truth label for the center pixel of each of the three locations per specimen (patient).
- Since this approach only allows a limited number of labeled lumpectomy data, it is unlikely that we can adequately train a supervised classification method, in particular a deep neural network, on the current dataset. On the other hand, the trained classifiers on the breast tissue slices can also not be directly used on the lumpectomy dataset. Although both the breast tissue slices and lumpectomy datasets have a similar domain, they differ in terms of tissue thickness, freshness, surface structure, blood saturation and cauterization. Hence, we can not expect a high performance from the currently developed tissue classifier, which is trained on slices and thus less suitable for analyzing the lumpectomy dataset. Kho *et al.* [28] reported a significantly lower performance when a LDA classifier, trained on a breast tissue slices dataset, was applied on a lumpectomy dataset.
- The ground truth (labels) at tissue transition areas are less reliable than on the breast tissue slices since the lumpectomy specimens are even more deformed during histopathological assessment which leads to registration inaccuracies of the H&E sections and HS images [29]. Besides that, the pixels at the transition areas may contain a mixture of different tissue types and thus the diffuse reflectance spectra of these pixels might not represent a single tissue type as is assumed by the ground truth labels.

To address the mentioned challenges, a neural network should be trained while exploiting both the data of breast tissue slices as a source domain and lumpectomy specimens with a sufficient number of malignant tumor as a target domain. One feasible way is to use domain adaptation (DA) which is a specific scenario of transfer learning. DA techniques use labeled data of a source domain for training a network that can be applied to classify the data of a target domain [30]. Due to differences between the domains, a network trained on the source domain likely has a lower performance on the target domain. Hence, over the past few years several methods, also for HS imaging classification in particular, have been proposed to overcome this problem [31]. Instance-based methods are applied in combination with active learning to iteratively select the most informative data with a query function to define a training set, as done by Tuia *et al.* [32]. The feature-based methods rely on feature extraction or feature selection. In [33,34], the authors used a transfer component analysis and canonical correlation analysis to extract features that minimize the differences between the domains. Therefore, the extracted features of both domains could be used by the same classifier.

In this work, we propose a deep learning method for the classification of HS images on breast lumpectomy specimens. The novel contributions of this paper can be summarized as follows:

- We developed one spectral and two spectral-spatial convolutional neural networks for the discrimination of healthy and tumor tissue in breast tissue slices, and to build a framework for the classification of lumpectomy specimens (section 2.2.5).
- We introduced a new loss function to account for label uncertainty at tissue transition areas where the ground truth labels were less reliable due to tissue deformations and mixtures of different tissue types (section 2.2.2). The effectivity on the classification performance is demonstrated in section 3.3.1.
- We acquired a labeled dataset of lumpectomy specimens that were measured immediately after surgery. By inking up to three locations on the lumpectomy surface we could make a direct correlation with histopathology, which allowed us to determine the ground truth labels for the measured locations.
- The main contribution of this paper is the introduction of a fine-tuning based domain adaptation approach to classify the HS images of breast lumpectomy specimens. This approach allowed us to retrain the previously developed neural networks on breast tissue slices and fine-tune the top layers with lumpectomy data so that we could predict the tissue percentages of the lumpectomy resection surface with HS imaging.

The remainder of this paper is organized as follows: Section 2 describes the data collection, data preparation and the proposed classification methods. Section 3 presents the experimental results, followed by the discussion and conclusion in Section 4 and 5, respectively.

## 2. Materials and methods

### 2.1. Materials

#### 2.1.1. Hyperspectral imaging setup

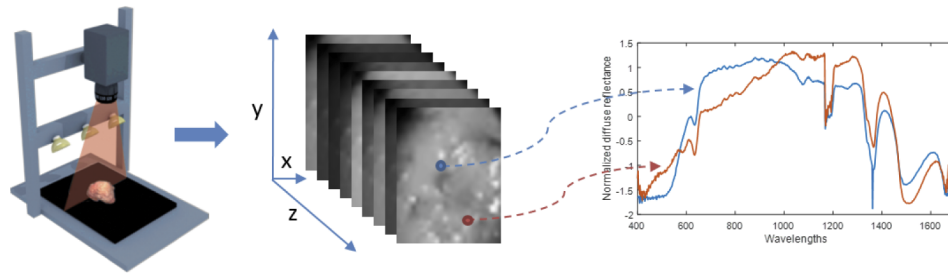
HS data were acquired with two pushbroom hyperspectral imaging cameras (Specim, Spectral Imaging Ltd., Finland) in the visible (VIS: PFD-CL-65-V10E, CMOS sensor  $1312 \times 384$  pixels,  $\sim 400$ - $1000$  nm, 384 wavelength bands, 3 nm increments) and near-infrared range (NIR: VLNIR CL-350-N17E, InGaAs sensor  $320 \times 256$  pixels  $\sim 900$ - $1700$  nm, 256 wavelength bands, 5 nm increments). The spatial resolutions were 0.16 mm/pixel and 0.5 mm/pixel for the VIS camera and NIR camera respectively. The tissue samples were illuminated by three halogen light sources (2900 K) mounted under identical 35-degree angles, and imaged line-by-line creating a 3D data structure or hypercube of which the first two dimensions represent the sample's spatial information and the third dimension the sample's spectral information, as shown in Fig. 2.

Data analysis was performed on the diffuse reflectance spectra of both systems, thus the raw data was first normalized into diffuse reflectance as described in [16]. Hereafter, the HS images of both systems were spatially matched using an affine transformation to resize the images and match the resolutions (0.5 mm/pixel). This resulted in one hyperspectral image with the size of  $320 \times 256$  pixels and 640 wavelength bands.

#### 2.1.2. Study design

The datasets were acquired on the excised breast tissue of female patients that had primary BCS in the Netherlands Cancer Institute - Antoni van Leeuwenhoek Hospital in the period from 2017 to 2020. These *ex vivo* studies were approved by the Institutional Review Board of the Netherlands Cancer Institute - Antoni van Leeuwenhoek hospital and complied with the Declaration of Helsinki. No written consent from the patients was required according to the Dutch medical research involving human subjects act (WMO).





**Fig. 2.** HS data acquisition and HS data. The tissue samples were imaged line-by-line with the HS imaging systems, creating a 3D data structure which is also called a hypercube. The x and y dimensions of this hypercube contain the spatial information, while each pixel contains one spectrum in the z dimension. Each pixel in spatial domain contains spectral information along all wavelengths as shown in the plot.

### 2.1.3. Data acquisition

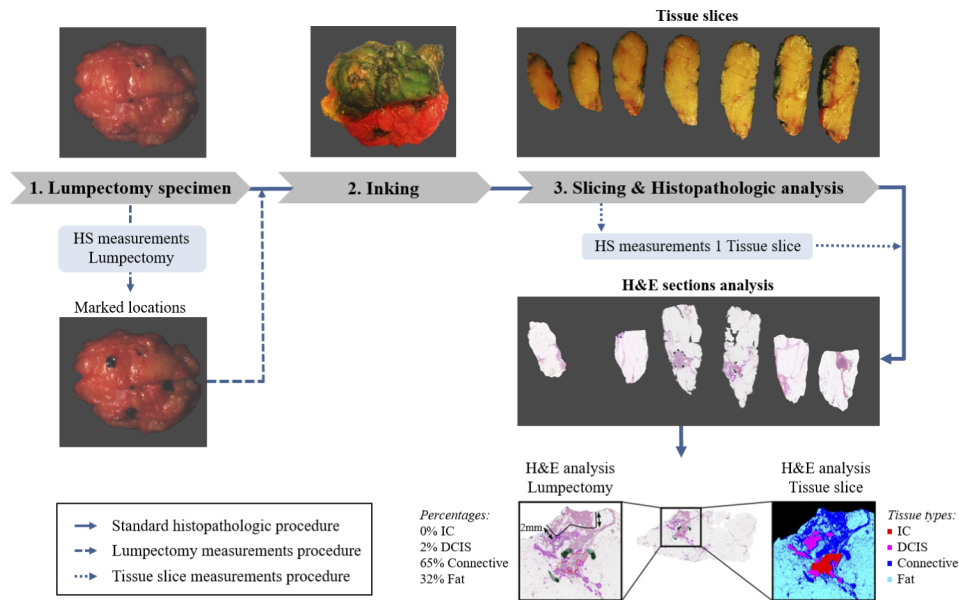
In this study, we used two datasets for the development and testing of the deep neural networks: a breast tissue slices dataset and a lumpectomy dataset. The acquisition and histopathology correlation of these datasets were performed in a similar approach as described in [16]. An overview of the acquisition method can be found in Fig. 3 and can be summarized as follows:

### 2.1.4. Breast tissue slices dataset

After surgery, the resected specimen was inked and gross-sectioned in tissue slices at the pathology department. To perform the optical measurements, one slice was selected which contained both healthy and tumor tissue. Both sides of this slice were imaged with the two HS imaging systems of which the total acquisition time included 2 minutes (each side: 40 seconds VIS camera and 20 seconds NIR camera). After the measurements, the tissue slice was processed into hematoxylin and eosin (H&E) stained sections and analyzed by a pathologist to annotate the surface with four tissue classes: invasive carcinoma (IC), ductal carcinoma in situ (DCIS), connective tissue, and fat tissue. The H&E section was registered with the HS image to determine the ground truth label for each pixel of the HS image.

### 2.1.5. Lumpectomy dataset

The lumpectomy specimen was measured immediately after surgery. First, the tissue was regarded as a cube with six resection sides. These sides were analyzed to select one side which was most likely to have a tumor-positive margin (only one side could be selected to obtain a high correlation with histopathology). This was performed by imaging the sides with initially the VIS camera only, to reduce the optical measurement time (each side: 40 seconds VIS camera and in total for six sides: 4 minutes), and subsequently classifying the images with a LDA algorithm that was previously trained on the breast tissue slices dataset [16,28]. The most suspicious side, according to the classifier, was selected and imaged again with both cameras (40 seconds VIS camera and 20 seconds NIR camera). This side was imaged two times: with and without black ink marks. The black ink marks are necessary to obtain the ground truth of the tissue but the ink affects the diffuse reflectance spectra. Therefore, also an image was taken prior to inking the locations. Subsequently, the lumpectomy specimen was processed according to standard procedure and the corresponding H&E sections, which contained the ink mark locations, were annotated with the aforementioned tissue classes (section 2.1.4) in a similar approach as described in [28]. Areas with IC and DCIS were annotated by a pathologist (Fig. 3) whereas areas with connective and fat tissue could be identified based on their color appearance in the H&E sections (i.e. pink



**Fig. 3.** Overview of the acquisition of the breast tissue slices dataset and lumpectomy dataset [28]. For the lumpectomy dataset, the specimen was imaged directly after surgery as described in Section 2.1.5. Subsequently, up to three locations on the lumpectomy surface were marked with black ink to enable correlation with histopathology. After data acquisition, the lumpectomy was further processed according to standard procedure including inking and slicing of the tissue. For the breast tissue slices dataset, the lumpectomy specimen was first inked and gross-sectioned in tissue slices at the pathology department. One tissue slice was selected for the hyperspectral measurements that consisted of both healthy and tumor tissue. Hereafter, along with the remaining tissue slices, this measured slice was further processed into hematoxylin and eosin (H&E) stained sections and analyzed by a pathologist. The H&E sections were used to obtain the ground truth of both datasets. For the lumpectomy dataset, the tissue up to 2 mm underneath the ink mark locations was analyzed using the H&E sections to obtain the percentage of invasive carcinoma (IC), ductal carcinoma in situ (DCIS), connective tissue and fat tissue. For the breast tissue slices dataset, the H&E sections were used to annotate the entire tissue slice as IC, DCIS, connective or fat tissue [28].

and white respectively). Hence, these tissue classes were annotated by thresholding the green channel of the corresponding H&E section at a value of 0.9. Lastly, the annotated H&E sections were used to determine the percentage of each tissue class up to 2 mm underneath the ink mark locations. These ground truth percentages corresponded to the center pixels of the black spots on the HS image.

## 2.2. Methodology

This subsection will first explain the steps that were taken to preprocess the HS images in order to deal with the noisy wavelengths, oblique illumination and rough tissue surface. Hereafter, a new loss function is introduced to account for label uncertainty in the breast tissue slices dataset, followed by a description of the networks used for the classification of the breast tissue slices dataset. At last, the domain adaptation methods will be explained that are used for the classification of the lumpectomy dataset. The classifications were performed in Python 3.7 on a machine with an NVIDIA GeForce GTX 1080 Ti.

### 2.2.1. Data preprocessing

Prior to tissue classification, HS images were first normalized into diffuse reflectance [16]. Both cameras have a low sensitivity at their spectral range's extremities. Therefore, these wavelengths were excluded from the analysis and only the wavelengths between 450-951 nm (318 wavelength bands) and 954-1650 nm (210 wavelength bands) were used for the VIS and NIR camera, respectively. The last step was standardizing the spectra using standard normal variate (SNV) to eliminate the spectral variability of each tissue type. The variability of the spectra was caused by the oblique illumination during scanning and the rough tissue surface [16].

### 2.2.2. Label uncertainty at tissue transition

The histopathology results (H&E sections) were registered with the HS images of the breast tissue slices dataset to annotate the HS images with the four tissue classes. These annotations were used as ground truth labels for each pixel of the breast tissue slices images. However, the labels at tissue transition areas are less reliable since the tissue will be deformed during the histopathological assessment which leads to registration inaccuracies of the H&E sections and HS images. Besides that, the pixels at the transition areas may contain a mixture of different tissue types and thus the diffuse reflectance spectra of these pixels might not represent a single tissue type as is assumed by the ground truth labels.

Therefore, we defined two different loss functions: 1) a Balanced Categorical Cross-Entropy (BCCE) loss function that includes all pixels, and 2) a Pixel Distance Excluding (PDE) loss function that excludes pixels based on their distance to the transition border to reduce the effect of label uncertainty.

The BCCE loss function is defined as:

$$L_{BCCE}(y, \hat{y}) = - \sum_{nc}^{NC} w_c y_{nc} \log(\hat{y}_{nc}) \quad (1)$$

where  $y_{nc}$  is the ground truth value and  $\hat{y}_{nc}$  is the output of the network for each pixel  $n$  of each class  $c$ . To correct for the imbalanced tissue types in the dataset, the weights  $w_c$  were included in the loss function and are defined as:

$$w_c = \frac{N}{C \sum_n y_{nc}} \quad (2)$$

where  $N$  is the number of pixels and  $C$  the total number of classes.



The proposed PDE loss function is a categorical cross-entropy loss including distance weights (DW) for each pixel and is defined as:

$$L_{PDE}(y, \hat{y}) = - \sum_{nc}^{NC} DW_n w_c y_{nc} \log(\hat{y}_{nc}) \quad (3)$$

A pixel was either excluded with a  $DW_n$  of 0 or included with a  $DW_n$  of 1. Pixels within a distance of 1 mm from the tissue transition border were recommended to be excluded by Kho *et al.* [16]. However, this results in a more imbalanced dataset containing a small number of DCIS pixels and a high number of fat pixels compared with IC and connective tissue. Therefore, the exclusion distance criterion was defined for each tissue type separately. For IC, DCIS, connective, and fat pixels with a distance from the tissue transition border lower or equal to 1 mm, 0.5 mm, 1 mm, and 3 mm respectively were excluded by the PDE loss function. By excluding pixels at the transition border of the HS image using the same distance criterion, a "CERTAIN" test set was prepared, while the "ALL" test set contained all pixels of the HS image.

### 2.2.3. Tissue classification in breast tissue slices using deep learning neural networks

In this study, we developed three different CNNs for the classification of breast tissue slices which included one spectral and two spectral-spatial networks. To enable a good comparison between the different networks, the breast tissue slices dataset was split into the same training set (55% of the patients), validation set (15% of the patients), and test set (30% of the patients) while keeping the images from one patient together. The training and validation sets were used to train the networks and to allow hyperparameter tuning, whereas the test set was used to evaluate the classification performance. In addition, a 5-fold cross-validation was performed to verify that the test set represented the whole dataset.

#### *1D-CNN-based spectral network*

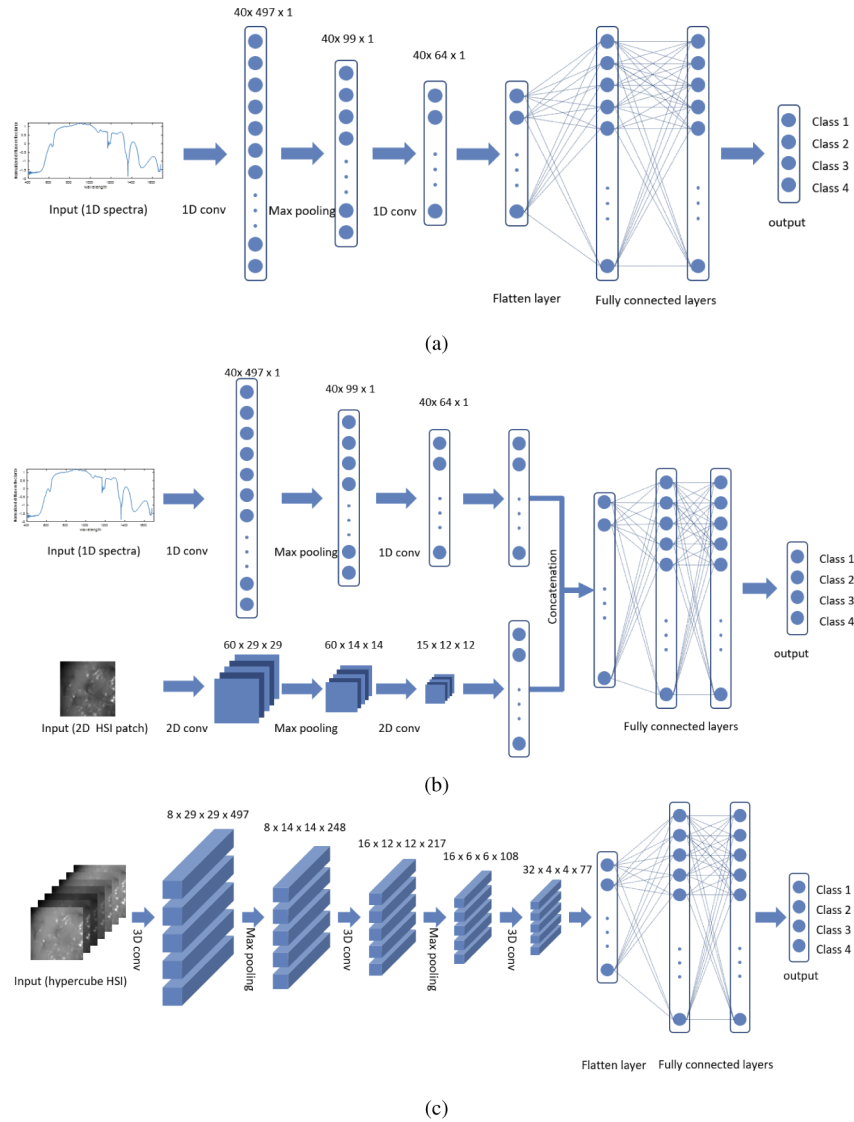
First, we developed a spectral network to classify the tissue types considering only their spectral information. Figure 4(a) shows the architecture of this 1D-CNN-based network. The SNV normalized spectrum  $sn$  of the  $n$ th pixel can directly be fed as input to the network. The model consists of three layers including 1D convolution to extract the spectral features followed by a nonlinear activation function (ReLU) and afterward a max-pooling operator to reduce the spectra size. The spectral features were flattened into a 1D feature vector and fed into the fully connected layers. A four-class softmax classifier was used to obtain the final classification results.

#### *Double-channel spectral-spatial network*

HS images contain, besides spectral information, also spatial information. Therefore, we developed a double-channel CNN (DC-CNN) similar to the one proposed by Yang *et al.* [26] which incorporates both spectral and spatial features for the classification. The DC-CNN includes two separated channels for the spectral and spatial feature extractions, as shown in Fig. 4(b). The spectral channel takes the SNV normalized spectrum  $sn$  of the  $n$ th pixel as input. This channel has the same layers as the 1D-CNN where the softmax classifier is replaced by a fully connected layer. After the convolution and max-pooling operations, the spectral features were obtained.

The spatial channel takes a patch  $P_n$  with neighboring pixels of the  $n$ th pixel as input. Before extracting the patch, the whole image was first preprocessed using principal component analysis (PCA) on the SNV normalized image. The patch size was fixed to  $31 \times 31$ . Several layers of 2D convolution including a ReLU activation function and max-pooling operations were applied to the spatial patch to extract the spatial features.

The network contains a high number of parameters due to the large size of the input which increases the likelihood of overfitting. Therefore, L2 regularization was applied to all convolutional layers. The spectral and spatial features were concatenated and simultaneously fed to two fully connected layers to extract joint-spectral-spatial features. The last fully connected layer was



**Fig. 4.** The architecture of neural networks for tissue classification. (a) the 1D-CNN-based spectral network, (b) Dual-channel network incorporating spectral and spatial information, (c) 3D-CNN-based spectral-spatial network.

followed by a four-class softmax classifier to predict the probability distribution for each tissue type.

#### 3D-CNN-based spectral-spatial network

The DC-CNN involves 1D and 2D convolutional layers to obtain the spectral and spatial features, respectively. However, 3D convolution can also be used to incorporate both spectral and spatial information. Therefore, we developed the 3D-CNN-based spectral-spatial network to extract the spectral and spatial features simultaneously. First, a hyperspectral patch  $HP_n$  with neighboring pixels of the  $n$ th pixel was extracted from the SNV normalized image. The patch size was fixed to  $31 \times 31 \times 528$ . Second, the input was fed to multiple layers of 3D convolution and max-pooling, as shown in Fig. 4(c). Dropout and L2 regularization were used to handle overfitting caused by the large number of parameters. At last, a four-class softmax layer was used for the final prediction.

#### 2.2.4. Training and hyperparameter tuning

During training, the BCCE or PDE loss functions (as described in Section 2.2.2) were minimized by optimizing the trainable parameters using stochastic gradient descent during 50 epochs. The learning rate and momentum were set to 0.001, 0.0009, 0.0003, and 0.95, 0.96, 0.98 for the 1D-CNN, DC-CNN, 3D-CNN, respectively.

Bayesian optimization was performed to tune the hyperparameters of the network. The hyperparameters include the variables that determine the network structure and how the network is trained (i.e. learning rate and momentum) [35]. A hypermodel has to be defined with the search space of the hyperparameters that need to be tuned. The search spaces based on the researches of Li *et al.* [23] and Chen *et al.* [19] are shown in Table 1. The Matthews Correlation Coefficient was chosen as the objective function since it can handle the imbalanced dataset. This will be further explained in Section 2.2.6.

**Table 1. Search range for hyperparameters tuning for 1D-CNN, DC-CNN and 3D-CNN networks**

Parameters	1D-CNN	DC-CNN	3D-CNN
Number of hidden layers	[1,2,3]	[1,2,3]	-
Number of filters per spectral conv. layers	[10,20,40]	[10,20,40]	[8,16,32]
Number of filters per spatial conv. layers		[15,30,60]	-
Number of fully connected layers	[0,1,2,3]	[1,2,3]	[1]
Number of neurons per fully connected layers	[200,400,600]	[200,400,600]	-
Learning rate	[1e-5 ; 1e-3]	[1e-5 ; 1e-3]	[1e-4 ; 1e-2]
Momentum	[0.9 ; 0.99]	[0.9 ; 0.99]	[0.9 ; 0.99]
Dropout rate in conv. layers	-	-	[0 ; 0.5]

First, a few combinations of hyperparameters were randomly chosen to evaluate the objective function. Second, A Gaussian process model was fitted through this observed data for approximating the objective function. Lastly, the acquisition function was used to determine the next combination of hyperparameters given the evaluation results. When the next combination was evaluated, the Gaussian process model was updated and the steps were repeated until the best model was found or the maximum of 100 trials was reached.

#### 2.2.5. Tissue classification in lumpectomy specimens using domain adaptation techniques

To use HS imaging as a margin assessment technique, a deep learning network should achieve a high classification performance on the lumpectomy dataset. However, obtaining sufficient labeled training samples in the lumpectomy dataset is challenging. Firstly, not every measured side of the resection surface contains tumorous tissue. Secondly, the ground truth cannot be obtained for the

whole surface since the histopathological margin assessment only covers a small fraction. As a result, the number of labeled training samples of the lumpectomy was insufficient for developing a reliable classification algorithm.

One solution is to use a domain adaptation strategy which applies the knowledge from the source domain to train a network for the target domain. Since the breast tissue slices dataset contains sufficient labeled data, we used this dataset as the source domain. The lumpectomy dataset was intended as the target domain. The lumpectomy dataset was split into a training, validation and test set using similar ratios as used for splitting the breast tissue slices dataset.

In this study, we used a fine-tuning based domain adaptation method to predict the tissue percentages of the lumpectomy resection surface. The bottom layers of this neural network extract low-level features, which are more generic than the high-level features, and could be transferred to the target domain. The top layers extract high-level features, thus those layers need to be trained on the target domain. First, the neural network was pre-trained on the breast tissue slices dataset using the PDE loss function defined in Eq. (3). Then, the top layers, i.e. fully connected layer and classification layer, were retrained during 100 epochs on the labeled lumpectomy dataset using the BCCE loss function defined in Eq. (1). Bayesian optimization was used to determine the optimal learning rate and momentum for the fine-tuning (1D-CNN: Learning rate=0.0002, Momentum=0.9; DC-CNN: Learning rate=0.0008, Momentum=0.9).

Since the lumpectomy locations contain a mixture of tissue types, fine-tuning was also performed with the percentages of each tissue type as ground truth label for each location rather than one single tissue type label. The learning rate and momentum were also tuned with Bayesian optimization (1D-CNN: Learning rate=0.001, Momentum=0.93; DC-CNN: Learning rate=0.0002, Momentum=0.99). The root mean square error was chosen as the objective function since the error between the ground truth percentages and predicted percentages should be minimized.

#### 2.2.6. Performance metrics and statistical analyses

The test sets of the breast tissue slices and lumpectomy specimens were used to evaluate the classification performance of the networks. Since it is clinically relevant to differentiate between healthy (connective tissue & fat) and tumor tissue (IC & DCIS), we evaluated the recall values to determine the percentages of pixels that were correctly classified as either tumor or healthy tissue. Furthermore, the Matthews Correlation Coefficient (MCC), sensitivity, specificity, and accuracy were calculated. The true positive (TP) rate was defined as the percentage of IC and DCIS pixels that were correctly classified as tumor tissue whereas the true negative (TN) rate was the percentage of connective and fat pixels correctly classified as healthy tissue. The false negative (FN) rate indicated the percentage of IC and DCIS pixels that were classified as healthy tissue, and the false positive (FP) rate the percentage of connective and fat pixels classified as tumor tissue. We used the MCC instead of the accuracy because it is regarded as a more robust metric for imbalanced sample sizes in the dataset. The MCC was calculated as follows:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (4)$$

The values range between -1 and 1, with 1 showing a perfect correlation, -1 no correlation, and 0 showing that the results were uncorrelated with the ground truth [36]. During training of the networks, this performance metric was evaluated for the validation set to determine the best model and to avoid overfitting.

We evaluated the predicted tissue percentages of the lumpectomy locations by calculating the root mean square error (RMSE). The RMSE represents the distance between the true percentages

on the H&E sections and the predicted percentages and can be calculated by

$$RMSE = \sqrt{\sum_c^C \sum_n^N \frac{(\hat{p}_{nc} - p_{cn})^2}{NC}} \quad (5)$$

where  $p_{nc}$  is the ground truth percentage and  $\hat{p}_{nc}$  is the predicted percentage by the network for each sample  $n$  of each class  $c$ .  $N$  and  $C$  represent the total number of samples and classes, respectively.

For comparison of the classification performance, the paired nonparametric McNemar's test was used to evaluate whether the networks differ significantly in terms of performance. The McNemar's test is based on a Chi-square statistics applied to a 2x2 contingency table. The null hypothesis states that the proportion of correctly classified pixels is similar for the two networks [37]. When the  $p$  value was smaller or equal to 0.05, the null hypothesis was rejected.

### 3. Results

#### 3.1. Data description

Tables 2 and 3 give an overview of the number of patients and pixels used per tissue type in the breast tissue slices and lumpectomy dataset. From these tables several differences can be observed: the breast tissue slices contain more labeled data than the lumpectomy specimens since the ground truth label could be obtained for each pixel in the HS image, whereas for the lumpectomy specimens the ground truth label could only be determined for up to three locations in the HS image (i.e. by using the black ink marks on the H&E sections). Thus, the remaining pixels were used as unlabeled data. In addition, for the breast tissue slices dataset, pixels that likely contained a mixture of different tissue types could be excluded by the PDE loss function during training. For the lumpectomy dataset, the three locations on each specimen had to be marked blindly before the H&E sections were obtained, so no information about the distribution of the tissue types was known. As a result, all marked locations in the lumpectomy dataset reflect a mixture of different tissue types. However, a location was labeled as tumor (i.e. IC or DCIS) if some percentage of tumor tissue was found underneath the black spot even if the percentage was lower than the healthy tissue. Therefore, no distinctive tissue labels were available for these marked locations.

**Table 2. Data description of the breast tissue slices**

Tissue class	Training set	Test set	
		ALL	CERTAIN
#patients (#pixels)			
IC	11 (7,616)	11 (6,792)	10 (2,936)
DCIS	24 (9,024)	6 (852)	1 (212)
Connective	29 (39,792)	13 (7,952)	6 (1,206)
Fat	29 (65,930)	13 (26,176)	7 (4,008)
Total	29 (122,362)	13 (41,772)	13 (8,362)

#### 3.2. Exclusion based on the distance to tissue transition

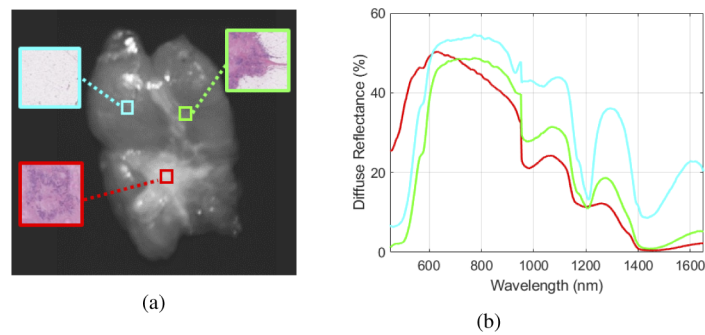
For the breast tissue slices dataset, the labels at tissue transition areas are less reliable. Figure 5 shows a representative example of a tissue transition and its corresponding spectrum: this spectrum (green) was neither equal to the spectrum taken in IC tissue (red) nor the spectrum taken in fat tissue (cyan). Instead, the spectrum represents a mixture of the two tissue types. Hence, to account for the label uncertainty, we defined the PDE loss function to exclude those



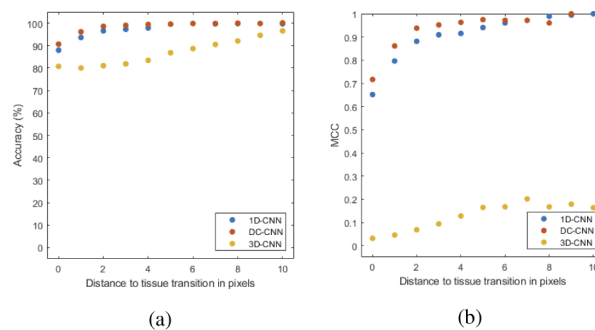
**Table 3. Data description of the lumpectomy specimens**

Tissue class	Training set		Test set
	labeled	Unlabeled	
#patients (#locations)			labeled
> 1% IC	18 (24)	-	5 (6)
> 1% DCIS	11 (14)	-	1 (2)
> 50% Connective	59 (94)	-	11 (17)
> 50% Fat	70 (94)	-	20 (32)
Total	96 (226)	96 (49,220)	25 (57)

pixels at tissue transitions from the training set. Figure 6 shows that a larger distance to the tissue transition increases the accuracy and MCC values for the pixels in the ALL test set.

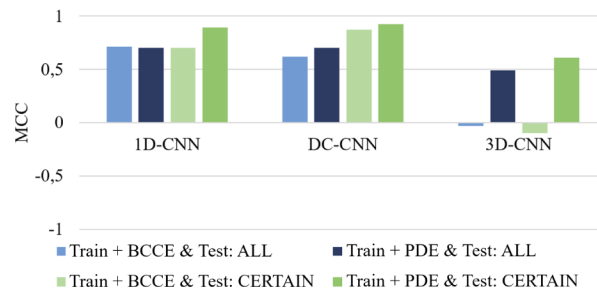


**Fig. 5.** Example of a tissue slice with an IC-fat tissue transition. The three locations (a) were taken in the middle of IC (red), in the middle of fat (cyan), and at the IC-fat tissue transition (green) as indicated by the corresponding H&E images. The three diffuse reflectance spectra (b) represent these locations.



**Fig. 6.** Classification performance vs distance to tissue transition. The accuracy in percentage (a) and MCC (b) with respect to the distance in pixels to the tissue transition for the 1D-CNN (blue dots), DC-CNN (red dots) and 3D-CNN (yellow dots).

After hyperparameter tuning, the networks were trained twice using the BCCE loss function including all samples as well as the PDE loss function to account for label uncertainty. Figure 7 shows the MCC values for all three networks evaluated on both the ALL and CERTAIN test sets. The highest MCC was achieved when the networks were trained with the PDE loss function and tested on the CERTAIN test set.



**Fig. 7.** Evaluation of the classification algorithms trained using the BCCE or PDE loss functions and tested on the ALL and CERTAIN dataset.

### 3.3. Classification results on breast tissue slices

In Tables 4 and 5 the performance metrics for the discrimination of tumor from healthy tissue and the recall values per tissue type are shown for each network respectively. The networks were trained twice using the BCCE loss function without accounting for label uncertainty, and the PDE loss function with accounting for label uncertainty.

**Table 4.** Performance metrics for the discrimination of tumor tissue from healthy tissue for 3 different networks trained using the BCCE and PDE loss functions

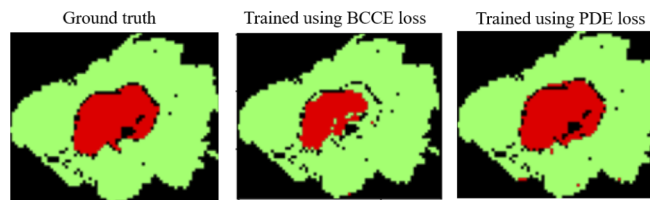
Test set		Training using BCCE Loss			Training using PDE loss		
		1D-CNN	DC-CNN	3D-CNN	1D-CNN	DC-CNN	3D-CNN
ALL	MCC	0.71	0.62	-0.03	0.70	0.70	0.49
	Sensitivity	0.67	0.62	0.00	0.87	0.78	0.72
	Specificity	0.97	0.95	0.36	0.90	0.94	0.84
	Accuracy	0.92	0.89	0.80	0.90	0.91	0.82
CERTAIN	MCC	0.70	0.87	-0.10	0.89	<b>0.92</b>	0.61
	Sensitivity	0.60	0.83	0.00	<b>0.91</b>	<b>0.91</b>	0.68
	Specificity	<b>1.00</b>	<b>1.00</b>	0.97	0.97	<b>1.00</b>	0.91
	Accuracy	0.85	0.94	0.61	0.95	<b>0.96</b>	0.82

**Table 5.** Classification results for 3 different networks Training using the BCCE or PDE loss function: recall values for each tissue type

Test set		Training using BCCE Loss			Training using PDE loss		
		1D-CNN	DC-CNN	3D-CNN	1D-CNN	DC-CNN	3D-CNN
ALL	IC	73%	64%	1%	92%	81%	70%
	DCIS	23%	51%	0%	46%	60%	83%
	Connective	95%	92%	98%	83%	87%	66%
	Fat	98%	96%	98%	93%	96%	89%
CERTAIN	IC	62%	84%	0%	94%	92%	66%
	DCIS	29%	67%	0%	45%	67%	97%
	Connective	100%	100%	100%	89%	100%	61%
	Fat	100%	100%	96%	100%	100%	100%

### 3.3.1. Accounting for label uncertainty improves tumor classification

In general, the classification performance (Table 4) improved when the algorithms were trained with the PDE loss function. The McNemar's tests show that the networks differ significantly in terms of performance when trained with the PDE loss function than with the BCCE loss function (1D-CNN:  $p < 0.0001$ ; DC-CNN:  $p < 0.05$ ; 3D-CNN:  $p < 0.02$ ). A higher sensitivity (Table 4) and recall for IC and DCIS (Table 5) were achieved with the PDE loss function. This is also illustrated in Fig. 8, which shows the classification results of the DC-CNN using the BCCE and PDE loss function for one breast tissue slice. In this slice more tumor pixels were correctly classified with the PDE loss function than with the BCCE loss function. However, with the BCCE loss function, a higher recall was achieved for connective tissue. The recall value for fat was similar for both loss functions.



**Fig. 8.** Classification results of tumor tissue (red pixels) and healthy tissue (green pixels) for one breast tissue slice using the DC-CNN. Left) the ground truth. Middle) classification results using the BCCE loss function (without accounting for label uncertainty). Right) classification results using the PDE loss function (with accounting for label certainty). The black pixels in the slice indicate pixels without a ground truth label.

### 3.3.2. DC-CNN outperforms 1D-CNN and 3D-CNN

For both loss functions the highest classification performance was achieved with the DC-CNN. This is shown in Table 4 where the performance metrics were either similar or higher compared to the 1D-CNN and 3D-CNN. Based on McNemar's test, the performance of the DC-CNN was significantly different from the other networks (1D-CNN:  $p < 0.0001$ ; 3D-CNN:  $p < 0.01$ ).

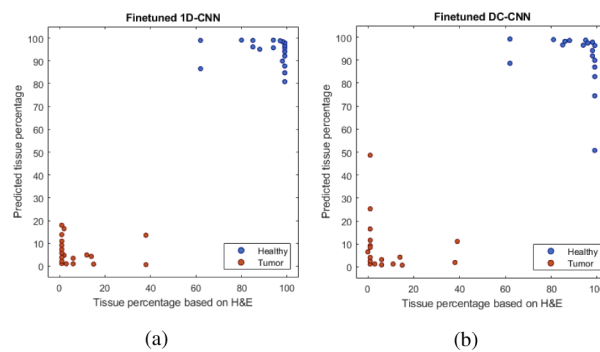
With regard to the PDE loss, differences were observed between the 1D-CNN and DC-CNN in the recall of DCIS and connective tissue. For both tissue classes, the DC-CNN achieved a higher recall (DCIS: 67%, Connective: 100%) than the 1D-CNN (DCIS: 45%, Connective: 89%). Comparing the DC-CNN and 3D-CNN, the recall values show that the DC-CNN was more capable of discriminating IC (92%) and connective tissue (100%) than the 3D-CNN (IC: 66%, connective: 61%). However, the 3D-CNN has a better performance on the discrimination of DCIS with a recall of 97%, whereas the DC-CNN achieved a recall of 67%. The recall value for fat was similar for all networks.

The results in Table 4 were compared using a single training and test set. Hence, 5-fold cross-validation was performed to confirm that the results represented the whole dataset. For the DC-CNN using the PDE loss function, the cross-validation shows a MCC, sensitivity and specificity range of respectively  $0.72 \pm 0.17$ ,  $0.82 \pm 0.13$  and  $0.89 \pm 0.14$ , which are thus comparable to the results in Table 4.

In summary, the results show that accounting for label uncertainty by using the PDE loss function improved the discrimination of tumor tissue (i.e. IC & DCIS) and healthy tissue. Adding spatial information to the classification algorithm using PCA (i.e. the DC-CNN) generally increased the classification performance.

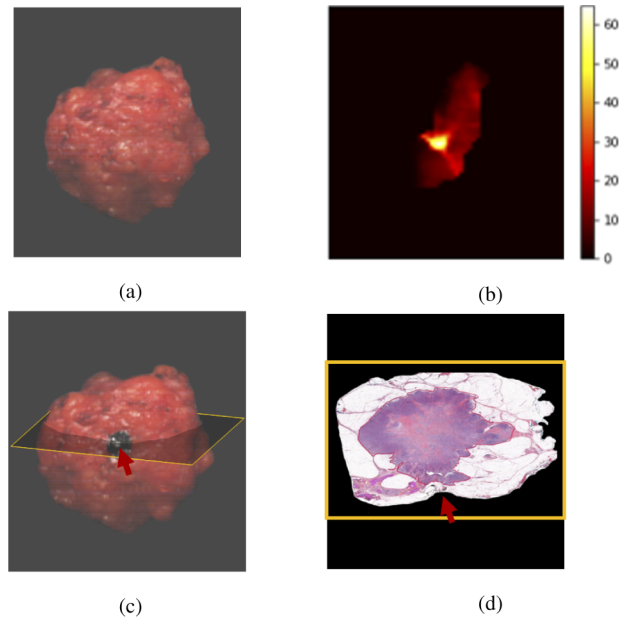
### 3.4. Classification results on lumpectomy dataset

The 1D-CNN and DC-CNN were fine-tuned to predict the tissue percentages in the lumpectomy specimens. Since the classification results of the 3D-CNN on the breast tissue slices were insufficient, this algorithm was excluded for the classification of the lumpectomy specimens. To fine-tune the networks, we used the 1D-CNN and DC-CNN that were trained with the PDE loss function on the breast tissue slices, and retrained the top layers with the BCCE loss function on the labeled lumpectomy training set. Subsequently, we evaluated the networks with the RMSE by using the labeled lumpectomy test set. The RMSE indicates the error of the predicted percentages of tumor and healthy tissue with respect to the ground truth percentages based on the H&E sections. The fine-tuned 1D-CNN and DC-CNN achieved an RMSE of respectively 9% and 11%. Figure 9 shows the predicted percentages of tumor and healthy tissue on the test set versus the ground truth percentages on the H&E sections.



**Fig. 9.** Predicted percentages by the 1D-CNN (a) and DC-CNN (b) vs. ground truth tissue percentages based on H&E sections of labeled lumpectomy locations. The tumor percentages are marked with red color and the healthy percentages with blue color.

These networks can be used to predict the tissue percentages for each pixel in the HS images of the lumpectomy. The prediction can support the surgeon in the decision to remove some additional tissue to ensure a complete excision. Figure 10 shows an example of the predicted tumor percentage for one side of the lumpectomy. In the middle, a high percentage of tumor tissue is predicted, so in this case the surgeon could decide to remove some extra tissue during surgery.



**Fig. 10.** Predicted tumor percentage of one side of the lumpectomy surface versus ground truth. In the middle of the lumpectomy surface (a) a tumor percentage of 60% is predicted (b). To verify this prediction with the ground truth, the suspected location (c) is marked with black ink (red arrow) and sliced (yellow rectangle). The corresponding ground truth cross-section (d) shows a tumor-positive area (red delineation) within 2 mm distance from the surface and thus confirms the tumor prediction.

## 4. Discussion

Achieving a tumor-negative resection margin is crucial to minimize the risk of tumor recurrence after breast-conserving surgery. However, this remains difficult as there is currently no margin assessment technique to provide real-time feedback during surgery. With the ability to image the entire resection surface in a rapid amount of time, HS imaging has the potential to overcome current limitations. Nevertheless, the development of an accurate deep classification network remains challenging because it is difficult to create a large dataset with ground truth labels on lumpectomy specimens.

In this study, we therefore used an extensive dataset on breast tissue slices to develop three convolutional neural networks for evaluating the classification performance of HS imaging. To make the step towards HS imaging analysis of lumpectomy specimens, we used a domain adaptation method to fine-tune these networks with lumpectomy data. Despite the use of a smaller labeled lumpectomy dataset, we were able to predict the tissue percentages on the lumpectomy resection surface.

### 4.1. Tissue classification in breast tissue slices

We expected a higher classification performance on the spectral-spatial networks since it was previously shown that adding spatial information to the network improved the classification performance [16,19]. Comparing the spectral 1D-CNN to the spectral-spatial DC-CNN, the performance of the DC-CNN was indeed higher than the 1D-CNN. However, we achieved a lower performance with the spectral-spatial 3D-CNN. This may be explained by the fact that the 3D-CNN contained a high number of trainable parameters, which increased the chance of



overfitting and generalization [38]. Since it is challenging to expand the data, several approaches were investigated to address the problem of overfitting: dropout, L2 regularization, and decreasing the number of layers. Nevertheless, the classification performance of the 3D-CNN remained lower than of the 1D-CNN and DC-CNN.

During the development of the classification algorithms on the breast tissue slices, we used the BCCE loss function to include all pixels as well as the PDE loss function to account for label uncertainty. Our results show that the highest classification performance was achieved when the networks were trained using the proposed PDE loss function. Besides the fact that excluding pixels with unreliable tissue labels improved the classification performance, also the use of hyperparameter tuning, with the PDE loss function rather than the BCCE loss function, resulted in a higher performance of the networks.

Using the PDE loss function, McNemar's test showed that the performance of the DC-CNN was significantly different from the 1D-CNN and the 3D-CNN. However, this test does not report which network performs significantly better. On top of that, McNemar's test does not measure the variability of the algorithms due to the choice of the training dataset. The algorithms were only compared using a single training and test set [39].

#### 4.2. *Tissue classification in lumpectomy specimens*

Domain adaptation techniques assume that the task is similar between the domains, while the domains are different. In this study, the task for both breast tissue slices and lumpectomy domain is similar since the HS images need to be classified with the same four tissue types. Concerning the domains, we expect that the feature space would be the same for both datasets since the same tissue types were imaged. However, we expect a spectral shift due to the difference in penetration depth of light, cutting method, and surface flatness which leads to different reflection values for each feature.

For the domain adaptation method, we used the PDE loss function for the breast tissue slices classification since this approach achieved the highest results on the CERTAIN test set. However, all locations in the lumpectomy dataset contain a mixture of different tissue types with no distinctive label. Therefore, it is possible that training using the BCCE loss function which includes all pixels of the breast tissue slices dataset improves the classification of the lumpectomy compared to the PDE loss function.

#### 4.3. *Limitations*

##### 4.3.1. Limited tumor in the lumpectomy data

In only 8 of the 57 locations of the test set, tumor tissue was found within 2 mm underneath the resection surface. These locations contained less than 40% tumor and thus, in this study the performance of the networks could not be evaluated for locations with higher tumor percentages. During data acquisition, the resection sides of the lumpectomy were evaluated using observation and palpation as well as a LDA classification algorithm to increase the likelihood of selecting tumor-positive locations. Nevertheless, the number of tumor locations remained insufficient to create a representative training and test set.

##### 4.3.2. No optimal threshold for tissue percentages

During labeling of the lumpectomy dataset, a location was labeled as tumor-positive if the ground truth percentage of tumor tissue was higher than 1%. However, no optimal threshold could be selected with regard to these percentages since a higher threshold increased the number of false negatives whereas a lower threshold increased the number of false positives. On top of that, these ground truth percentages were based on a region of 2 mm underneath the resection surface. Since definitions on positive resection margins vary per country [40], the predicted tissue percentages should only be considered as an approach to identifying tumor suspicious areas rather than

detecting positive resection margins. Nevertheless, they can support the surgeon in the decision to remove some additional tissue to ensure a complete excision.

#### 4.4. Comparison to similar work

##### 4.4.1. Classification methods

This is the first study in which a domain adaptation method was used to improve the tissue classification of breast lumpectomy HS images. However, in literature different methodologies are used for classifying HS images of breast tissue specimens [16,28,41–43].

Panasyuk *et al.* [41] were one of the first pioneers who performed an in vivo HS imaging study on 56 rats with induced mammary tumors. Different from our study, the authors used a liquid crystal tunable filter-based HS system in the visual range (450–700 nm, 34 wavelength bands) to acquire the data of 339 individual sample locations and developed a classification algorithm on the corresponding absorption spectra. Several tissue types such as tumor, muscle, connective tissue (including fat) and even blood vessels could be clearly detected on the classified HS images. On top of that, they achieved a performance of 89% sensitivity and 94% specificity to detect residual tumor tissue using histopathology as the gold standard. Since these results were only demonstrated in animal models, no definite conclusions on human subjects could be drawn. Therefore, it is not possible to make a comparison with our study.

In a study of Pourreza-Shahri *et al.* [42] the authors performed an ex vivo study on 19 human breast tissue specimens. A digital light processing-based HS system was used between 380 and 780 nm at 101 different wavelengths. Subsequently, the most important features were extracted with a Fourier Coefficient Selection features approach followed by a Minimum Redundancy Maximum Relevance method to reduce spectral dimensionality. Hereafter, the authors used a SVM classifier with a radial basis kernel function to distinguish healthy from tumor tissue with a sensitivity of 98% and specificity of 99%. Despite these high performance results, this study only evaluated tumor, fat and connective tissue whereas we also included DCIS in the classification because this type of tissue significantly adds to the number of positive resection margins [3]. Since DCIS is the precursor of tumor tissue (IC), these small premalignant cells are usually difficult to detect with HS imaging [16]. Hence, this explains a lower sensitivity (91%) on the test set.

Similar to our study, Aboughaleb *et al.* [43] used a pushbroom HS system to discriminate ex vivo healthy from tumor tissue in human breast specimens. However, the data were acquired over a smaller wavelength range between 420 and 620 nm and six bands. For the classification of the ten included specimens, the authors applied a moving average filter and subsequently a K-mean clustering algorithm. A sensitivity and specificity of respectively 95% and 96% were obtained, which are in line with our results. Nevertheless, there is a major difference with our study: per patient the tumor and healthy data were acquired on different tissue samples whereas we obtained the data on the same sample. In other words, the authors used the resected breast specimens as tumor tissue and removed another part of the breast (at 5–10 cm distance from the tumor) to use it as healthy tissue. These samples consisted of pure tissue only while we particularly included mixtures of tissue.

Compared to the previous publications, Kho *et al.* [16] acquired one of the most extensive datasets on breast tissue slices. This dataset consisted of 42 patients with over 300.000 spectra, and was obtained with two pushbroom HS systems in both the visual and near-infrared range (450–1650 nm, 528 bands). By using a Fisher's Linear Discriminant Analysis classifier, the authors could discriminate healthy (connective tissue, fat) from tumor tissue (invasive carcinoma, ductal carcinoma in situ) with a very high performance of 98% sensitivity and 99% specificity. In a follow-up study [28], the authors made the step towards classification of lumpectomy specimens but due to a small lumpectomy dataset, they were not able to develop a new classifier. Therefore, the authors examined whether they could directly apply the classification algorithm that was

developed on the breast tissue slices, to the data of the lumpectomy specimens but found this to be insufficient for obtaining adequate classification results.

#### 4.4.2. Margin assessment techniques

Comparing our results on the lumpectomy specimens with the performances of other margin assessment techniques (Section 1.), the classification performance of HS imaging has been on the low side.

For ultrasound imaging, a sensitivity of 86% and specificity of 100% were reported [7]. Although promising, the performance of this technique mainly depends on the operator who should be highly experienced given that the outcome of the images is prone to interpretation errors. With HS imaging there is no need for an experienced operator as the performance rather depends on the outcome of the classification algorithm.

For optical coherence tomography, the reported sensitivity and specificity were 100% and 82% respectively [11], which outperforms our results on the lumpectomy specimens. Nevertheless, this technique has a small field-of-view that only covers approximately a region of 1 cm<sup>2</sup>, making it less effective than HS imaging to image the entire resection surface quickly.

Also frozen section analysis has high sensitivity and specificity of 83% and 95% respectively [6,44]. However, this method adds on average 27 minutes to the operation time. With HS imaging, the entire resection surface can be analyzed in less than two minutes including 1 minute for imaging the lumpectomy with the HS cameras, 20 seconds for preprocessing and 1 second for classifying the image with the DC-CNN. The classification performance of HS imaging on the breast tissue slices was comparable with frozen section analysis which shows potential for the classification of the lumpectomy specimens when more tumor data is available.

Besides HS imaging, there are also other spectroscopy methods (e.g. Raman spectroscopy) that have the potential of detecting cancer cells with a high performance [8–10]. However, in contrast to HS imaging, most of these methods only allow single-point measurements. Hence, to analyze the entire resection surface, multiple sites have to be measured which makes these methods rather laborious and time-consuming for use during surgery.

### 4.5. Future research

In future research, multiple directions can be pursued. In short, the results on the lumpectomy dataset can be improved by extending the amount of tumor data or using a different methodology for analyzing the data.

#### 4.5.1. Data acquisition

To improve the results, more data should be available on the lumpectomy specimen with a reliable tissue label. Training and testing of the network on spectra with a representative label should increase the classification performance. Therefore, more data should be acquired on locations containing a higher percentage of tumor tissue than healthy tissue. It might also be helpful to redefine the percentages and tissue labels for the H&E section with tumor delineations including healthy tissue.

#### 4.5.2. Methodological improvements

The lumpectomy dataset contains more unlabeled data compared with labeled data. Therefore, the performance could be improved by exploiting this unlabeled data using weakly supervised or unsupervised learning methods.

Thereby, the method of determining the tumor percentages should incorporate the penetration depth of light. One disadvantage of HS imaging is that the penetration depth varies both with wavelength and tissue types, thus it is not similar for all diffuse reflectance spectra. Since the percentages in this study were calculated for a fixed depth, the obtained spectra might not represent

the same area underneath the tissue surface due to a different penetration depth. Consequently, the tissue labels for the lumpectomy specimen might be inaccurate when not accounting for the penetration depth of light for each wavelength. This can be solved by either performing wavelength selection to ensure that only wavelengths with the desired penetration depth are included [45], or by training a deep learning model to recognize different penetration depths based on the intensity of the reflection spectra [46].

Furthermore, since the optical resolution of the light is lower than the spatial resolution of the HS cameras, it is also important to consider that the obtained spectrum from one pixel covers a larger sampling volume than solely the size of the pixel. In other words, one pixel might represent a mixture of surrounding tissue types rather than a single tissue type. In our study, we accounted for this by using tissue percentages instead of distinctive labels, and including neighboring pixels as input for the DC-CNN and 3D-CNN. However, hyperspectral unmixing could be another potential solution to distinguish the spectra from each other when no pure pixels exist [47].

## 5. Conclusion

In this study, we have demonstrated that the resection margins of breast lumpectomy specimens can be classified with HS imaging through a domain adaptation approach. Our results showed that the classification performance of the algorithm can be improved by exploiting HS images of both the breast tissue slices and the lumpectomy datasets using the proposed PDE loss function to account for label uncertainty. In particular, the discrimination of connective and fat tissue from tumor tissue was improved. Since the data represented a mixture of different tissue types, the lumpectomy resection surface was predicted with tissue percentages rather than distinctive labels. The prediction of the tissue percentages on the lumpectomy resection surface shows potential as an RMSE of 9% was achieved with the fine-tuned 1D-CNN. For further improvements, the lumpectomy dataset should be either expanded with locations that contain more tumor than healthy tissue or more reliable tissue percentages should be calculated based on the H&E sections and penetration depth of light.

**Funding.** KWF Kankerbestrijding (10747).

**Acknowledgments.** The authors thank the NKI-AVL core Facility Molecular Pathology & Biobanking (CFMPB) for supplying NKI-AVL biobank material, all surgeons and nurses from the Department of Surgery and all pathologist and pathologist assistants from the Department of Pathology for their assistance in collecting the specimens. The Quadro P6000 GPU used for this research was donated by the NVIDIA Corporation.

**Institutional Review Board.** This study was conducted at The Netherlands Cancer Institute (NKI-AVL) under approval of the Institutional Review Board (CFMPB545). This study was performed in compliance with the Declaration of Helsinki and approved by the Institutional Review Board of The Netherlands Cancer Institute/Antoni van Leeuwenhoek (Amsterdam, the Netherlands). According to Dutch law (WMO), no written informed consent from patients was required.

**Disclosures.** The authors declare no conflicts of interest.

**Data availability.** Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

## References

1. H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *Ca-Cancer J. Clin.* **71**(3), 209–249 (2021).
2. L. E. McCahill, R. M. Single, E. J. Aiello Bowles, H. S. Feigelson, T. A. James, T. Barney, J. M. Engel, and A. A. Onitilo, "Variability in reexcision following breast conservation surgery," *JAMA - J. Am. Med. Assoc.* **307**(5), 467–475 (2012).
3. C. A. Garcia-Etienne, M. Tomatis, J. Heil, K. Friedrichs, R. Kreienberg, A. Denk, M. Kiechle, F. Lorenz-Salehi, R. Kimmig, G. Emons, M. Danaei, V. Heyl, U. Heindrichs, C. J. Rageth, W. Janni, L. Marotti, M. R. Del Turco, A. Ponti, L. Cataliotti, E. Cretella, P. Van Dam, A. Emons, T. Gyr, R. Hils, P. Kern, U. Koehler, S. Kuemmel, D. Liedtke, A. Luini, V. Moebus, M. Neumann, S. Paepke, O. Pagani, L. Pavesi, D. Sarlos, T. Schlotfeldt, C. Sohn, A. Spelsberg, G. Staelens, M. Taffurelli, C. Tinterri, I. B. Vergin, T. Zemmler, and D. Wagner, "Mastectomy trends for early-stage breast cancer: a report from the EUSOMA multi-institutional European database," *Eur. J. Cancer* **48**(13), 1947–1956 (2012).

4. M. Morrow, K. J. Van Zee, L. J. Solin, N. Houssami, M. Chavez-MacGregor, J. R. Harris, J. Horton, S. Hwang, P. L. Johnson, M. L. Marinovich, S. J. Schnitt, I. Wapnir, and M. S. Moran, "Society of Surgical Oncology–American Society for Radiation Oncology–American Society of Clinical Oncology Consensus Guideline on Margins for Breast-Conserving Surgery With Whole-Breast Irradiation in Ductal Carcinoma in Situ," *Pract. Radiat. Oncol.* **6**(5), 287–295 (2016).
5. D. E. Wazer, T. DiPetrillo, R. Schmidt-Ullrich, L. Weld, T. J. Smith, D. J. Marchant, and N. J. Robert, "Factors influencing cosmetic outcome and complication risk after conservative surgery and radiotherapy for early-stage breast carcinoma," *J. Clin. Oncol.* **10**(3), 356–363 (1992).
6. J. J. Keating, C. Fisher, R. Batiste, and S. Singhal, "Advances in intraoperative margin assessment for breast cancer," *Curr. Surg. Rep.* **4**(4), 15 (2016).
7. T. E. Doyle, R. E. Factor, C. L. Ellefson, K. M. Sorensen, B. J. Ambrose, J. B. Goodrich, V. P. Hart, S. C. Jensen, H. Patel, and L. A. Neumayer, "High-frequency ultrasound for intraoperative margin assessments in breast conservation surgery: A feasibility study," *BMC Cancer* **11**(1), 444 (2011).
8. I. Pappo, R. Spector, A. Schindel, S. Morgenstern, J. Sandbank, L. T. Leider, S. Schneebaum, S. Lelcuk, and T. Karni, "Diagnostic performance of a novel device for real-time margin assessment in lumpectomy specimens," *J. Surg. Res.* **160**, 277 (2010).
9. A. S. Haka, Z. Volynskaya, J. A. Gardecki, J. Nazemi, R. Shenk, N. Wang, R. R. Dasari, M. Fitzmaurice, and M. S. Feld, "Diagnosing breast cancer using Raman spectroscopy: prospective analysis," *J. Biomed. Opt.* **14**(5), 054023 (2009).
10. S. Dhar, J. Y. Lo, G. M. Palmer, M. A. Brooke, B. S. Nichols, B. Yu, N. Ramanujam, and N. M. Jokerst, "A diffuse reflectance spectral imaging system for tumor margin assessment using custom annular photodiode arrays," *Biomed. Opt. Express* **3**(12), 3211 (2012).
11. F. T. Nguyen, A. M. Zysk, E. J. Chaney, J. G. Kotynek, U. J. Oliphant, F. J. Bellafiore, K. M. Rowland, P. A. Johnson, and S. A. Boppart, "Intraoperative evaluation of breast tumor margins with optical coherence tomography," *Cancer Res.* **69**(22), 8790–8796 (2009).
12. G. Lu and B. Fei, "Medical hyperspectral imaging: a review," *J. Biomed. Opt.* **19**(1), 010901 (2014).
13. M. Halicek, J. D. Dormer, J. V. Little, A. Y. Chen, L. Myers, B. D. Sumer, and B. Fei, "Hyperspectral imaging of head and neck squamous cell carcinoma for cancer margin detection in surgical specimens from 102 patients using deep learning," *Cancers* **11**(9), 1367 (2019).
14. B. Fei, G. Lu, X. Wang, H. Zhang, J. V. Little, M. R. Patel, C. C. Griffith, M. W. El-Diery, and A. Y. Chen, "Label-free reflectance hyperspectral imaging for tumor margin assessment: a pilot study on surgical specimens of cancer patients," *J. Biomed. Opt.* **22**(08), 1 (2017).
15. F. Manni, R. Fonollà, F. van der Sommen, S. Zinger, C. Shan, E. Kho, S. B. de Koning, T. Ruers, and P. H. de With, "Hyperspectral imaging for colon cancer classification in surgical specimens: towards optical biopsy during image-guided surgery," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (IEEE, 2020), pp. 1169–1173.
16. E. Kho, B. Dasthozorg, L. L. de Boer, K. K. Van de Vijver, H. J. C. M. Sterenborg, and T. J. M. Ruers, "Broadband hyperspectral imaging for breast tumor detection using spectral and spatial information," *Biomed. Opt. Express* **10**(9), 4496 (2019).
17. H. Petersson, D. Gustafsson, and D. Bergström, "Hyperspectral image analysis using deep learning - A review," in *2016 6th International Conference on Image Processing Theory, Tools and Applications, IPTA 2016* (Institute of Electrical and Electronics Engineers Inc., 2017).
18. Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature* **521**, 436 (2015).
19. Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sensing* **54**(10), 6232–6251 (2016).
20. Q. Gao, S. Lim, and X. Jia, "Hyperspectral image classification using convolutional neural networks and multiple feature learning," *Remote Sens.* **10**(2), 299 (2018).
21. X. Yang, Y. Ye, X. Li, R. Y. Lau, X. Zhang, and X. Huang, "Hyperspectral image classification with deep learning models," *IEEE Trans. Geosci. Remote Sensing* **56**(9), 5408–5423 (2018).
22. Y. Luo, J. Zou, C. Yao, X. Zhao, T. Li, and G. Bai, "HSI-CNN: a novel convolution neural network for hyperspectral image," in *ICALIP 2018 - 6th International Conference on Audio, Language and Image Processing* (2018), pp. 464–469.
23. Y. Li, H. Zhang, and Q. Shen, "Spectral–spatial classification of hyperspectral imagery with 3D vconvolutional neural network," *Remote Sens.* **9**(1), 67 (2017).
24. W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, "Deep convolutional neural networks for hyperspectral image classification," *J. Sens.* **2015**, 1–12 (2015).
25. S. Yu, S. Jia, and C. Xu, "Convolutional neural networks for hyperspectral image classification," *Neurocomputing* **219**, 88–98 (2017).
26. J. Yang, Y. Q. Zhao, and J. C. W. Chan, "Learning and transferring deep joint spectral-spatial features for hyperspectral classification," *IEEE Trans. Geosci. Remote Sensing* **55**(8), 4729–4742 (2017).
27. H. Zhang, Y. Li, Y. Zhang, and Q. Shen, "Spectral-spatial classification of hyperspectral imagery using a dual-channel convolutional neural network," *Remote. Sens. Lett.* **8**(5), 438–447 (2017).



28. E. Kho, B. Dashtbozorg, J. Sanders, M.-J. T. Vrancken Peeters, F. van Duijnhoven, H. J. Sterenborg, and T. J. Ruers, "Feasibility of ex vivo margin assessment with hyperspectral imaging during breast-conserving surgery: From imaging tissue slices to imaging lumpectomy specimen," *Appl. Sci.* **11**(19), 8881 (2021).
29. L. L. de Boer, E. Kho, J. Nijkamp, K. K. Van de Vijver, H. J. Sterenborg, L. C. ter Beek, and T. J. Ruers, "Method for coregistration of optical measurements of breast tissue with histopathology: the importance of accounting for tissue deformations," *J. Biomed. Opt.* **24**(07), 1 (2019).
30. W. M. Kouw and M. Loog, "A review of domain adaptation without target labels," Tech. rep. (2019).
31. D. Tuia, C. Persello, and L. Bruzzone, "Domain adaptation for the classification of remote sensing data: An overview of recent advances," *IEEE Geosci. Remote Sens. Mag.* **4**, 41 (2016).
32. D. Tuia, E. Pasolli, and W. J. Emery, "Using active learning to adapt remote sensing image classifiers," *Remote. Sens. Environ.* **115**(9), 2232–2242 (2011).
33. G. Matasci, M. Volpi, M. Kanevski, L. Bruzzone, and D. Tuia, "Semisupervised Transfer Component Analysis for Domain Adaptation in Remote Sensing Image Classification," *IEEE Trans. Geosci. Remote Sensing* **53**(7), 3550–3564 (2015).
34. A. A. Nielsen, "The regularized iteratively reweighted MAD method for change detection in multi- and hyperspectral data," *IEEE Trans. on Image Process.* **16**(2), 463–478 (2007).
35. J. Prost, "Determine Your Network Hyperparameters With Bayesian Optimization," Sicara
36. S. Boughorbel, F. Jarray, and M. El-Anbari, "Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric," *PLoS One* **12**(6), e0177678 (2017).
37. J. de Leeuw, H. Jia, L. Yang, X. Liu, K. Schmidt, and A. K. Skidmore, "Comparing accuracy assessments to infer superiority of image classification methods," *Int. J. Remote. Sens.* **27**(1), 223–232 (2006).
38. S. Prasad and J. Chanussot, *Hyperspectral Image Analysis: Advances in Machine Learning and Signal Processing*, Advances in Computer Vision and Pattern Recognition (Springer International Publishing).
39. T. G. Dietterich, "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms," Tech. Rep. (1997).
40. S. G. B. de Koning, M.-J. T. Vrancken Peeters, K. Józwiak, P. A. Bhariosing, and T. J. Ruers, "Tumor resection margin definitions in breast-conserving surgery: systematic review and meta-analysis of the current literature," *Clin. Breast Cancer* **18**(4), e595–e600 (2018).
41. S. V. Panasyuk, S. Yang, D. V. Faller, D. Ngo, R. A. Lew, J. E. Freeman, and A. E. Rogers, "Medical hyperspectral imaging to facilitate residual tumor identification during surgery," *Cancer Bio. Therapy* **6**(3), 439–446 (2007).
42. R. Pourreza-Shahri, F. Saki, N. Kehtarnavaz, P. Leboulluec, and H. Liu, "Classification of ex-vivo breast cancer positive margins measured by hyperspectral imaging," in *2013 IEEE International Conference on Image Processing* (IEEE, 2013), pp. 1408–1412.
43. I. H. Aboughaleb, M. H. Aref, and Y. H. El-Sharkawy, "Hyperspectral imaging for diagnosis and detection of ex-vivo breast cancer," *Photodiagn. Photodyn. Ther.* **31**, 101922 (2020).
44. K. Esbona, Z. Li, L. G. Wilke, L. G. Wilke, and A. Surg, "Intraoperative imprint cytology and frozen section pathology for margin assessment in breast conservation surgery: a systematic review," *Ann. Surg. Oncol.* **19**(10), 3236–3245 (2012).
45. E. Kho, L. L. de Boer, A. L. Post, K. K. Van de Vijver, K. Józwiak, H. J. Sterenborg, and T. J. Ruers, "Imaging depth variations in hyperspectral imaging: Development of a method to detect tumor up to the required tumor-free margin width," *J. Biophotonics* **12**(11), e201900086 (2019).
46. F. Geldof, B. Dashtbozorg, B. H. Hendriks, H. J. Sterenborg, and T. J. Ruers, "Layer thickness prediction and tissue classification in two-layered tissue structures using diffuse reflectance spectroscopy," *Sci. Rep.* **12**(1), 1698 (2022).
47. B. Rasti, B. Koirala, P. Scheunders, and J. Chanussot, "Miscinet: Minimum simplex convolutional network for deep hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sensing* **60**, 5522815 (2022).