



# Supporting intraoperative margin assessment using deep learning for automatic tumour segmentation in breast lumpectomy micro-PET-CT



Luna Maris<sup>1,2</sup>✉, Menekse Göker<sup>3</sup>, Kathia De Man<sup>4</sup>, Bliede Van den Broeck<sup>4</sup>, Sofie Van Hoecke<sup>5</sup>, Koen Van de Vijver<sup>6,7</sup>, Christian Vanhove<sup>1,8</sup> & Vincent Keereman<sup>1,2</sup>

Complete tumour removal is vital in curative breast cancer (BCa) surgery to prevent recurrence. Recently, [<sup>18</sup>F]FDG micro-PET-CT of lumpectomy specimens has shown promise for intraoperative margin assessment (IMA). To aid interpretation, we trained a 2D Residual U-Net to delineate invasive carcinoma of no special type in micro-PET-CT lumpectomy images. We collected 53 BCa lamella images from 19 patients with true histopathology-defined tumour segmentations. Group five-fold cross-validation yielded a dice similarity coefficient of  $0.71 \pm 0.20$  for segmentation. Afterwards, an ensemble model was generated to segment tumours and predict margin status. Comparing predicted and true histopathological margin status in a separate set of 31 micro-PET-CT lumpectomy images of 31 patients achieved an F1 score of 84%, closely matching the mean performance of seven physicians who manually interpreted the same images. This model represents an important step towards a decision-support system that enhances micro-PET-CT-based IMA in BCa, facilitating its clinical adoption.

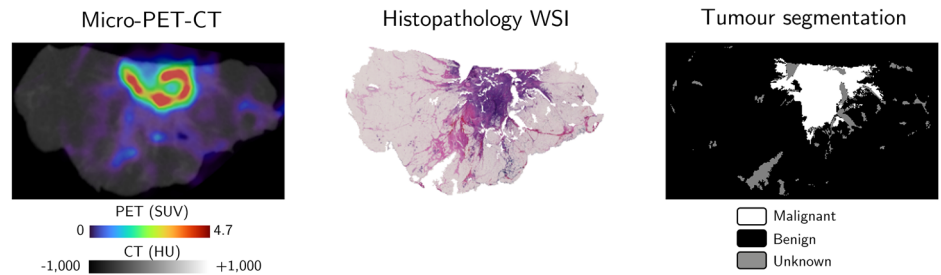
Early-stage breast cancer (BCa) is commonly treated with breast-conserving surgery (BCS), followed by radiotherapy<sup>1,2</sup>. BCS, also called lumpectomy, involves local surgical excision of the tumour. The procedure is considered successful when the primary tumour is completely removed, along with a margin of surrounding healthy tissue<sup>1</sup>. To assess whether the excision was adequate, the gold standard is sectional histopathology<sup>3</sup>. This process involves marking the surface of the resected specimen with ink to assess the margin status. According to the 2014 SSO-ASTRO consensus guidelines, a positive margin for invasive BCa is defined as “ink on tumour”, meaning the tumour reaches the surface of the excised specimen<sup>4</sup>. However, histopathological margin assessment requires significant time, typically taking at least 16 h, and often even 3–5 days<sup>3</sup>. Following initial excision, positive margins are found in around 20% of patients with invasive BCa, with reported rates ranging from 9% to 36%<sup>1,5</sup>. Since positive margins are the most important risk factor for local recurrence after BCS, patients with

inadequate margins often have to undergo re-surgery to remove any remaining tumour cells from the breast<sup>6</sup>.

To avoid re-surgery, intraoperative margin assessment (IMA) is crucial, allowing immediate removal of additional tissue in case of suspected positive surgical margins. Recently, specimen imaging with high-resolution positron emission tomography (micro-PET), in combination with high-resolution computed tomography (micro-CT), emerged as a promising technique for IMA during BCS<sup>7–9</sup>. For this technique, the radiotracer [<sup>18</sup>F] fluorodeoxyglucose (FDG) is preoperatively administered to the patient. Afterwards, during BCS, a dedicated micro-PET-CT specimen scanner is used to visualise the resected lumpectomy specimen in 3D, allowing fast assessment of the resection margins. Pilot studies where physicians interpreted micro-PET-CT images of lumpectomy specimens in terms of their margin status, were published by Göker et al.<sup>7</sup> and De Crem et al.<sup>9</sup>. The most recent study is the one by De Crem et al.<sup>9</sup> where seven physicians, with varying levels of experience, postoperatively assessed the margin status of

<sup>1</sup>Ghent University, Department of Electronics and Information Systems, MEDISIP, Ghent, Belgium. <sup>2</sup>XEOS Medical, R&D Department, Ghent, Belgium. <sup>3</sup>Ghent University Hospital, Department of Gynaecology, Ghent, Belgium. <sup>4</sup>Ghent University Hospital, Department of Medical Imaging, Nuclear Medicine, Ghent, Belgium. <sup>5</sup>Ghent University-Imec, Internet Technology and Data Science Lab (IDLab), Ghent, Belgium. <sup>6</sup>Ghent University Hospital, Department of Pathology, Ghent, Belgium. <sup>7</sup>Ghent University, Department of Diagnostic Sciences and CRIG, Ghent, Belgium. <sup>8</sup>Ghent University, CORE ARTH INFINITY, Ghent, Belgium. ✉e-mail: [luna.maris@ugent.be](mailto:luna.maris@ugent.be)

**Fig. 1 | Example of a micro-PET-CT slice with the corresponding true tumour segmentation, as annotated on the whole slide image (WSI) by the pathologist, for an invasive carcinoma of no special type (NST). On the left, the 2D micro-PET-CT image, aligned with the histopathology WSI, is visualised. The micro-PET and micro-CT images are expressed in standardised uptake values (SUVs) and Hounsfield units (HUs), respectively. In the middle, the WSI is shown. On the right, the binary tumour annotation is visualised. Malignant (NST), benign, and unknown pixels are shown in white, black, and grey, respectively.**



lumpectomy specimens based on their micro-PET-CT images. Specimens included 31 invasive carcinomas of no special type (NST), 6 invasive lobular carcinomas (ILCs), and 2 cases of ductal carcinoma in situ (DCIS). Comparing the predicted to the true histopathological margin status resulted in a mean sensitivity of 90% and a mean specificity of 86% for the NST cases. For the ILC and DCIS cases, the average sensitivities were 50% and 100%, and the specificities were 54% and 100%, respectively. While it was found that the specificity increased for the NST and ILC cases with increasing experience level of the physicians, the seven physicians showed substantial to almost perfect agreement with Fleiss' kappa scores of 0.80, 0.62, and 1.00 for the NST, ILC, and DCIS cases, respectively<sup>9</sup>.

Although these results are promising for NST and DCIS, we hypothesise that a decision-support system, to aid physicians in micro-PET-CT interpretation, might further enhance inter-reader reliability, reduce the time needed to interpret the images intraoperatively, and perhaps even improve sensitivity and specificity. The need for automatic (artificial intelligence based) algorithms that can enhance the interpretation of micro-PET-CT images, was previously also expressed by Lambert et al.<sup>8</sup> and Moraitis et al.<sup>10</sup>. In this work, we present a deep learning (DL) model that automatically segments (pre)malignant breast tissue in micro-PET-CT lumpectomy images, thereby also enabling automatic margin status prediction. Such DL model can form the basis of a clinical decision-support tool.

## Results

### Data sets

To train and test the DL model, we collected two data sets with different types of ground truth labels, derived from histopathology. The first data set contains 2D micro-PET-CT slices of lumpectomy lamellas together with their true histopathological semantic tumour annotations. This data set was used for supervised training and testing of the DL model for tumour segmentation. The second data set contains 3D micro-PET-CT images of entire lumpectomy specimens for which the true histopathological margin status (positive or negative) is known. This data set was used to validate the DL model in terms of margin assessment accuracy.

For the first data set with tumour segmentation labels, we collected micro-PET-CT images of lumpectomy lamellas from 28 BCa patients, including 19 NST, 6 ILC, and 3 DCIS cases. In total, we collected micro-PET-CT images of 53, 19, and 10 lamellas of the NST, ILC, and DCIS specimens, respectively, along with the corresponding histopathology whole slide images (WSIs). All micro-PET-CT images were co-registered with the WSIs. The NST, ILC, and DCIS tumours were annotated in the WSIs to obtain binary tumour segmentations, distinguishing between (pre)malignant and benign pixels. Figure 1 visualises an example of a 2D micro-PET-CT slice with the co-registered histopathology WSI and the corresponding true tumour annotation. For some tissue pixels in the micro-PET-CT images, the labels were unknown. The fraction of tissue pixels with an unknown label was on average  $11.12\% \pm 5.80\%$ ,  $8.27\% \pm 3.85\%$ , and  $21.92\% \pm 12.93\%$  across all patients with NST, ILC, and DCIS BCa, respectively. The data was also strongly imbalanced. The fraction of labelled tissue pixels that was (pre)malignant was on average  $16.66\% \pm 9.91\%$ ,

$9.55\% \pm 5.85\%$ , and  $0.34\% \pm 0.17\%$  across all patients with NST, ILC, and DCIS BCa, respectively.

For the second data set with margin classification labels, we collected 3D micro-PET-CT images of 39 BCa specimens, all from different patients, including 31 NST, 6 ILC, and 2 DCIS cases. Through histopathological analysis, 9 NST, 2 ILC, and 1 DCIS case were found to have a positive margin following the “ink on tumour” definition. This margin status is further indicated as the “true histopathological margin status” of the specimen. The maximal tumour diameter measured during histopathological processing of the specimens was on average  $18.19 \pm 9.07$  mm,  $20.17 \pm 8.03$  mm, and  $25.00 \pm 11.00$  mm for the NST, ILC, and DCIS cases. This data set was previously also described by De Crem et al.<sup>9</sup>. There is an overlap between both data sets used in this work: for three patients we collected both lamella images with the true tumour segmentation, as well as images of the full specimens with the true histopathological margin status. As explained later on, it was assured that models were never trained and validated on data from the same patient, avoiding data leakage.

### Trained models

Table 1 provides an overview of all models trained in this work. Using the first data set with true segmentation labels, we trained a 2D Residual U-Net (ResU-Net) to segment (pre)malignant breast tissue in micro-PET-CT images of lumpectomy specimens. A stratified group nested five-fold cross-validation scheme resulted in a total of 25 trained ResU-Nets, five per test fold. Models were trained for three different inputs: micro-CT, micro-PET, and micro-PET-CT. The same five-fold cross-validation scheme was used for the three different model inputs, resulting in a total of 75 trained models, 25 per model input type. All models were first trained using only the NST data, and transfer learning was performed for the ILC and DCIS data jointly. This doubled the amount of models, resulting in a final total of 150 models. An overview of the hyperparameters of the different models, optimised through grid search, can be found in Supplementary Table 1. It took around 15 min to train one ResU-Net on the NST data. Performing transfer learning for the combined ILC and DCIS cases took around 5 min per ResU-Net. Supplementary Figures 1 and 2 show the evolution of the train and validation loss across epochs for model training on the NST and combined ILC and DCIS cases, respectively.

We compared the performance of the ResU-Nets to a conventional intensity thresholding approach. We used the same stratified group five-fold train-test split to find the most optimal intensity thresholds to segment the (pre)malignant breast tissue in micro-PET-CT images. As the cross-validation scheme was not nested in this case, we obtained 5 intensity thresholds in total, one per test fold. Separate thresholds were defined for the three different model inputs (micro-CT, micro-PET, and micro-PET-CT) and the different tumour types (the NST tumours, and the joint ILC and DCIS tumours), resulting in a final total of 30 intensity thresholds.

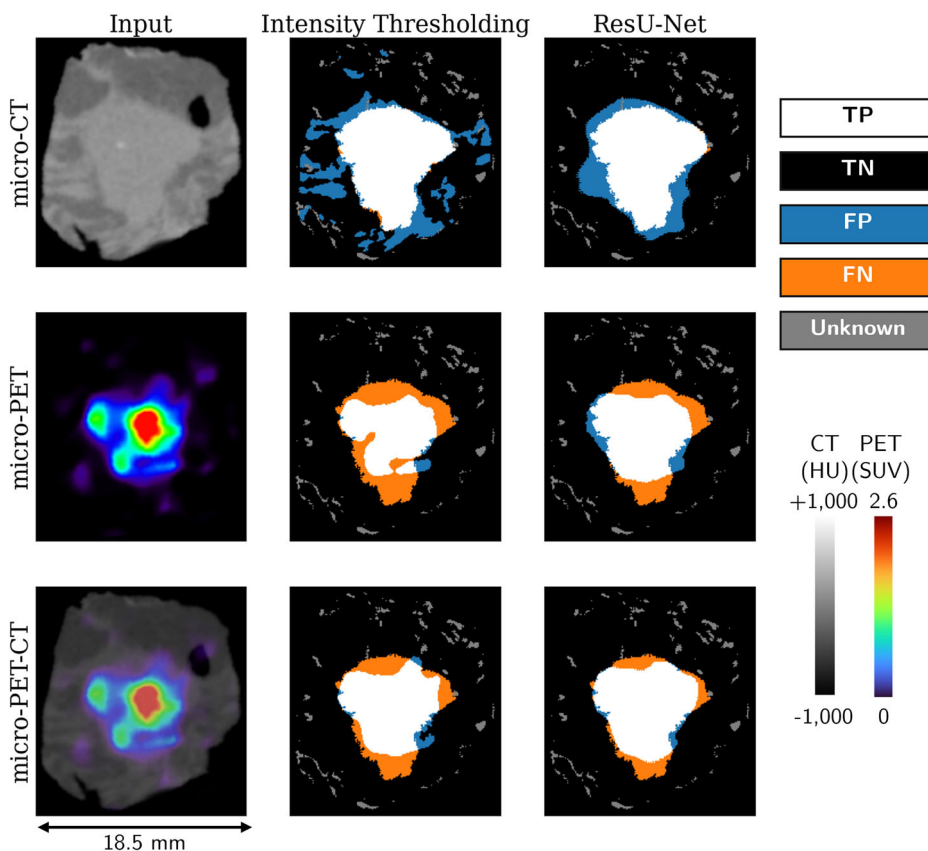
### Model performance evaluation: tumour segmentation

The trained models were evaluated in two ways. First, we evaluated the ability of the ResU-Nets and intensity thresholds to correctly segment (pre)

**Table 1 | Overview of all 150 Residual U-Nets (ResU-Nets) and 30 intensity thresholds (Intens thresh) trained in this work, with their respective input images (micro-CT, micro-PET, or micro-PET-CT), and the tumour type they were trained to segment (invasive carcinoma of no special type (NST), or invasive lobular carcinoma (ILC) and ductal carcinoma in situ (DCIS))**

Model input	Tumour type	# ResU-Nets	# ResU-Nets/test fold	# Intens thresh	# Intens thresh/test fold
micro-CT	NST	25	5	5	1
micro-PET	NST	25	5	5	1
micro-PET-CT	NST	25	5	5	1
micro-CT	ILC + DCIS	25	5	5	1
micro-PET	ILC + DCIS	25	5	5	1
micro-PET-CT	ILC + DCIS	25	5	5	1

**Fig. 2 | Example of the tumour segmentations predicted with intensity thresholding and Residual U-Net (ResU-Net) for an invasive carcinoma of no special type.** Predictions are shown for three different inputs: micro-CT, micro-PET, and micro-PET-CT. The micro-PET and micro-CT images are expressed in standardised uptake values (SUVs) and Hounsfield units (HUs), respectively. True positive (TP), true negative (TN), false positive (FP), and false negative (FN) pixels are shown in white, black, blue, and orange, respectively. Grey pixels indicate pixels with unknown labels.



malignant breast tissue. The trained ResU-Nets and intensity thresholds were used to predict the tumour segmentations for all test cases of the first data set with true segmentation labels. We made predictions for three model inputs: (1) micro-CT, (2) micro-PET, and (3) micro-PET-CT. For NST cases, tumour segmentations were predicted using the models trained only on NST data, while for ILC and DCIS cases, tumour segmentations were predicted using the models obtained after transfer learning on the joint ILC and DCIS data. Predicting the tumour segmentation for one 2D micro-PET-CT slice, using the appropriate ensemble of five trained ResU-Nets of the according fold, took less than one second. Figure 2 visualises the predicted tumour segmentations for a representative NST test case. By comparing the predicted and true tumour segmentations, we could calculate the dice similarity coefficient (DSC), the area under the curve of the precision-recall curve (PR-AUC), the 95th percentile of the Hausdorff distance (95HD), and the contour dice with a 1 mm tolerance (1mmCD). Table 2 shows the mean scores of the evaluation metrics across the test patients with NST tumours.

The tumour segmentation performance of the ResU-Nets and intensity thresholding models was compared using the Wilcoxon signed-rank test. Table 2 shows the *p*-values resulting from this analysis. The box plot in Figure 3 shows the distribution of the metric values in more detail for the NST tumours. Full test results for ILC and DCIS tumours can be found in the Supplementary Material. Given the limited number of data samples for these BCa types, the results should be interpreted with caution. Visualisations of predicted tumour segmentations for a representative ILC and DCIS case can be found in Supplementary Figures 3 and 4, respectively. The mean metric values across the test patients with ILC and DCIS tumours can be found in Supplementary Table 2. In addition, Supplementary Table 2 also shows the *p*-values resulting from the Wilcoxon signed-rank test to compare the performance of the ResU-Nets with the intensity thresholding models. The box plots in Supplementary Figures 5 and 6 show the distribution of the metric values in more detail for the ILC and DCIS tumours, respectively. Furthermore, the Wilcoxon signed-rank test was also conducted to compare

**Table 2 | Metric values to evaluate the agreement between true and predicted tumour segmentations for 19 test patients with invasive carcinoma of no special type (NST)**

Input:	micro-CT			micro-PET			micro-PET-CT		
BCa type:	NST (N = 19)								
Model:	Intens thresh	ResU-Net	<i>p</i>	Intens thresh	ResU-Net	<i>p</i>	Intens thresh	ResU-Net	<i>p</i>
DSC	0.58 ± 0.16	0.61 ± 0.17	0.2101	0.67 ± 0.22	0.69 ± 0.19	0.0955	0.71 ± 0.22	0.73 ± 0.20	0.1564
PR-AUC	0.59 ± 0.16	0.71 ± 0.19	<b>0.0028</b>	0.75 ± 0.21	0.79 ± 0.16	0.0874	0.77 ± 0.20	0.83 ± 0.14	<b>0.0000</b>
95HD	5.34 ± 2.56	4.01 ± 2.06	0.0230	3.36 ± 1.71	3.43 ± 1.60	0.7086	3.27 ± 1.78	3.15 ± 1.62	0.7086
1mmCD	0.40 ± 0.19	0.46 ± 0.18	0.2413	0.67 ± 0.22	0.61 ± 0.22	<b>0.0095</b>	0.69 ± 0.21	0.71 ± 0.19	0.3124

Mean and standard deviation across the test patients are shown for the following metrics: the dice similarity coefficient (DSC), the area under the curve of the precision-recall curve (PR-AUC), the 95th percentile of the Hausdorff distance (95HD) in mm, and the contour dice with a 1 mm tolerance (1mmCD). All metrics are shown per model input (micro-CT, micro-PET, or micro-PET-CT), and per segmentation strategy (Intensity thresholding (Intens thresh) or Residual U-Net (ResU-Net)). *P*-values (*p*) resulting from the Wilcoxon signed-rank test, comparing the performance of the ResU-Nets with the intensity thresholding models, are reported for each metric, stratified per model input. Significant performance differences following Bonferroni-Holm correction ( $\alpha = 0.05$ ,  $m = 4$ ) are indicated in bold.

the performance of the ResU-Nets and intensity thresholding models across different model inputs: micro-PET, micro-CT, and micro-PET-CT. The *p*-values resulting from this analysis are shown in Supplementary Table 3.

When looking at model performance for the NST tumours, best performance was found when combining micro-PET with micro-CT to predict tumour segmentations. For both the ResU-Nets and intensity thresholding models, using combined micro-PET-CT input generally resulted in significantly improved performance in terms of DSC and PR-AUC, compared to using micro-PET or micro-CT alone. For the NST cases, ResU-Net performed slightly better than intensity thresholding, but differences were small. For the models with micro-PET-CT input, significant differences were seen in the PR-AUC, where ResU-Net demonstrated greater precision than intensity thresholding and provided more consistency across test patients. While in general the same trends were seen for the ILC and DCIS cases, we could not show any significant differences in model performance. Tumour segmentation is particularly challenging for the DCIS cases, where we found a low performance in terms of DSC and PR-AUC. The 95HD scores showed large variability for all BCa types due to outliers with large contour distances. Most outliers arise from the model mistakenly identifying small tumour spots alongside the main tumour mass. A possible solution is to retain only the largest connected structure of the predicted tumour segmentation. However, this should only be applied in 3D to avoid unintended removal of parts of the main tumour mass.

### Model performance evaluation: margin assessment

Second, we evaluated whether the ResU-Nets and intensity thresholds could be used to correctly predict the margin status for 3D micro-PET-CT lumpectomy images. We used the appropriate ensemble of 25 trained ResU-Nets or 5 intensity thresholds to predict the tumour segmentation and margin status for the BCa specimens of the second data set with known true histopathological margin status. For this analysis, we only predicted tumour segmentations using the combined micro-PET-CT input, as this previously delivered the best results. Using the ensemble of 25 trained ResU-Nets, it took around 1 min in total to predict the tumour segmentation for a 3D micro-PET-CT image, slice-per-slice in three orthogonal directions. During this analysis, it was ensured that the models used to predict the tumour segmentation for a specimen were not trained on lamella data from the same specimen. As a consequence, for three tumour specimens, the prediction was made using only five out of 25 ResU-Nets, and one out of five intensity thresholds. Figure 4 visualises the tumour contours predicted with ResU-Net and intensity thresholding for two NST specimens. Visualisations of predicted tumour contours for two ILC and two DCIS specimens can be found in Supplementary Figures 7 and 8, respectively.

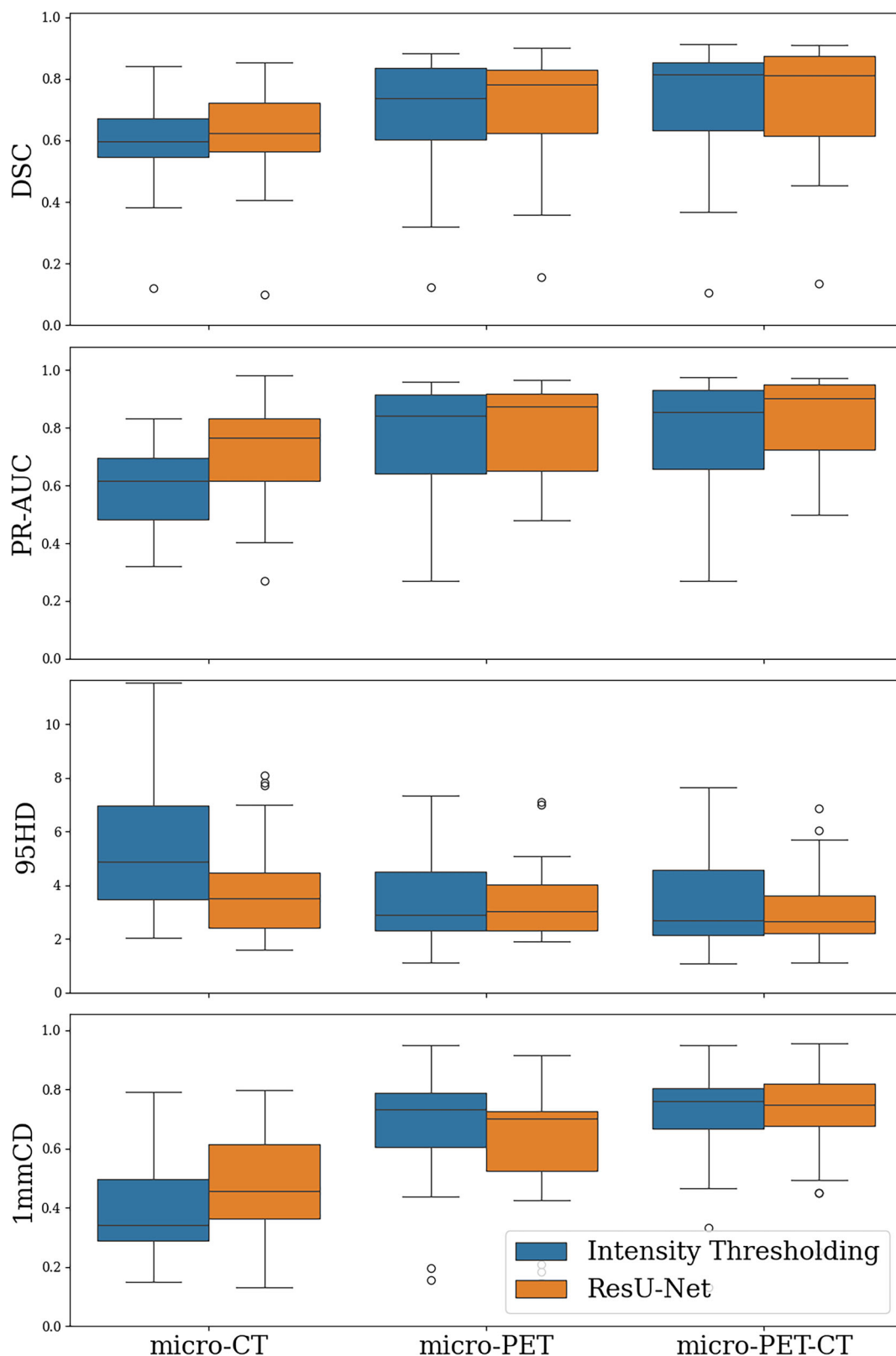
The predicted margin status was considered positive in case a minimal distance  $\leq 0.1$  mm was found between the specimen contour and the predicted tumour contour. We calculated the sensitivity, specificity, precision, and F1 score for positive margin detection by comparing the predicted and true histopathological margin status of all specimens. Table 3 shows all

metric values for the NST cases. The scores for the ILC and DCIS tumours are shown in Supplementary Table 4. The best results were obtained when only considering the largest connected structure of the predicted tumour segmentation. While this post-processing step entails a risk of missing tumour spots, it increased the specificity, while maintaining the same sensitivity for positive margin detection. When multifocal breast tumours are identified through preoperative imaging, this post-processing step can be disabled. However, multifocal BCa is often considered as a contraindication for BCS, making lumpectomy specimens with multifocal tumours uncommon<sup>11</sup>.

The scores are compared to the scores obtained by seven physicians who manually interpreted the same micro-PET-CT images. Table 3 and Supplementary Table 4 report two sets of evaluation scores: the scores obtained when using the majority vote of the seven physicians to define the margin status per BCa specimen, as well as the average scores reached across the physicians. These scores were previously described by De Crem et al.<sup>9</sup> *P*-values resulting from McNemar's test to evaluate differences in margin assessment performance between the ResU-Nets, intensity thresholding models, and the majority vote of seven physicians, are reported in Table 4 for the NST cases, and in Supplementary Table 5 for the ILC and DCIS cases. For the NST cases, the margin status predictions obtained with the ResU-Nets were the same as the ones obtained through majority voting of seven physicians. The intensity thresholding models showed significantly lower performance for margin assessment in the NST cases, with lower values for the specificity, precision, and F1 score. For the ILC cases, the highest performance scores were found through majority voting of seven physicians, followed by the ResU-Nets, and intensity thresholding models. For the DCIS tumours, the margin status was correctly predicted for the two available specimens using all three methods. Both for the ILC and DCIS cases, no significant difference in margin assessment performance between the three strategies could be shown.

### Discussion

To our knowledge, we were the first to apply the well-known ResU-Net DL model to perform automatic tumour segmentation and margin assessment based on micro-PET-CT images of lumpectomy specimens. First, we used a data set with true segmentation labels to train and evaluate a ResU-Net to segment (pre)malignant breast tissue in 2D micro-PET-CT lumpectomy images. The ResU-Net performed best when taking the combined micro-PET-CT images as input, compared to only providing one of both modalities to the model. This indicates that the combination of the metabolic information from micro-PET, with the higher spatial resolution from micro-CT, is crucial to accurately segment the tumour. The best performing ResU-Net reached a DSC of  $0.72 \pm 0.20$ , and a PR-AUC of  $0.83 \pm 0.14$  for the segmentation of invasive carcinoma of the NST type. While the performance of the ResU-Net was slightly higher compared to conventional intensity thresholding, the difference in NST segmentation performance was only found to be significant in terms of the PR-AUC. Second, we used a

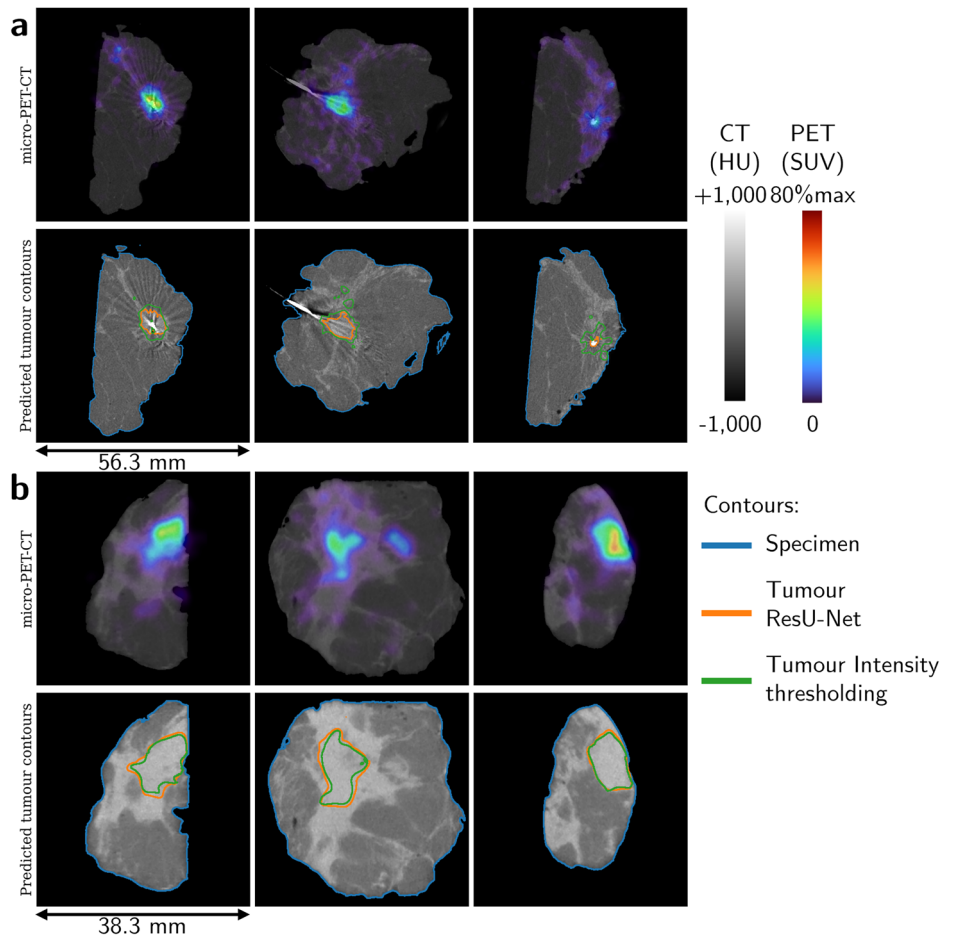


**Fig. 3 |** Box plot showing metric distributions assessing agreement between true and predicted tumour segmentations for test patients with invasive carcinoma of no special type. The following metrics are shown: the dice similarity coefficient (DSC), the area under the curve of the precision-recall curve (PR-AUC), the 95th percentile of the Hausdorff distance (95HD) in mm, and the contour dice with a 1 mm tolerance (1mmCD). The metrics are shown per model input (micro-CT, micro-PET, or micro-PET-CT) and per segmentation strategy (Intensity thresholding, or Residual U-Net (ResU-Net)).

**Fig. 4 | Predicted tumour contours for two specimens with invasive carcinomas of no special type.**

For every specimen, three orthogonal slices are visualised. The top row shows the micro-PET-CT input, and the bottom row shows the contours on top of the micro-CT. The micro-PET and micro-CT images are expressed in standardised uptake values (SUVs) and Hounsfield units (HUs), respectively. The specimen contour is shown in blue, and the contours of the tumour predictions are shown in orange and green for the Residual U-Net (ResU-Net) and intensity thresholding, respectively.

**a** Specimen with negative histopathological margin. The ResU-Net prediction is true negative, while intensity thresholding leads to a false positive prediction. **b** Specimen with positive histopathological margin. Both ResU-Net and intensity thresholding predictions are true positive.



**Table 3 | Metric values to evaluate the agreement between the true histopathological and predicted margin status for 31 invasive carcinomas of no special type (NST)**

BCa type:	Intens thresh	ResU-Net	7 Physicians (Majority vote)	7 Physicians (Average)
NST (N = 31)				
Sensitivity (Recall)	1.00	0.89	0.89	0.90 ± 0.04
Specificity	0.59	0.91	0.91	0.86 ± 0.07
Precision	0.50	0.80	0.80	0.74 ± 0.08
F1 score	0.67	0.84	0.84	0.81 ± 0.04

The following metrics are shown: sensitivity (or recall), specificity, precision, and F1 score. Margin status was predicted using intensity thresholding (Intens thresh) and the Residual U-Net (ResU-Net) applied to the micro-PET-CT input. The performance of both methods is compared to the performance of seven physicians manually interpreting the same images. Both the scores obtained through a majority vote of these seven physicians, as well as the mean scores (± standard deviation) across the seven physicians are shown. The performance scores for these seven physicians were previously described by De Crem et al.<sup>9</sup>.

data set of 3D micro-PET-CT lumpectomy images with true histopathological margin status labels to evaluate the performance of the ResU-Net in terms of margin assessment. For NST tumours, the ResU-Net reached an F1 score of 84%, matching the majority vote of seven physicians, and outperforming intensity thresholding. While similar trends were observed for the ILC and DCIS tumours, the limited data sets for these cases did not allow to show any significant differences in performance between the ResU-Nets and intensity thresholding models for tumour segmentation, nor for margin assessment.

Other groups have used DL to create similar workflows for automatic IMA in BCa based on other intraoperative imaging techniques: Veluponnar

**Table 4 | P-values (p) resulting from McNemar’s test to evaluate differences when predicting the margin status for 31 invasive carcinomas of no special type (NST) with the following three strategies: majority voting of seven physicians, prediction with Residual U-Nets (ResU-Nets), and prediction with intensity thresholding (Intens thresh) models**

BCa type	Margin assessment strategy	p
NST (N = 31)	7 Physicians (Majority vote) vs. ResU-Net	1.0000
	7 Physicians (Majority vote) vs. Intens thresh	<b>0.0078</b>
	ResU-Net vs. Intens thresh	<b>0.0078</b>

Significant performance differences following Bonferroni-Holm correction ( $\alpha = 0.05, m = 3$ ) are indicated in bold.

et al.<sup>12</sup> and Yeung et al.<sup>13</sup> used ultrasound (US) images, Shia et al.<sup>14</sup> used tomosynthesis images, and Kho et al.<sup>15</sup> used hyperspectral images of lumpectomy specimens. All four groups first developed DL models to segment tumours in their respective lumpectomy images. Veluponnar et al.<sup>12</sup> reported a median DSC of 0.88 on a set of 109 US images of 86 patients, and Yeung et al.<sup>13</sup> reported a mean DSC of  $0.96 \pm 0.05$  on a test set of around 732 US images of three patients. Shia et al.<sup>14</sup> reported a mean DSC of 0.86 on a test set of five breast tomosynthesis images of five patients. Kho et al.<sup>15</sup> did not report a DSC, but they reached a pixel-level sensitivity of  $0.76 \pm 0.24$  and precision of  $0.93 \pm 0.09$  for a test set of 26 hyperspectral images of 13 tissue slices of 13 patients. Although we obtained a lower DSC of  $0.73 \pm 0.20$  for the NST tumours, the scores are not directly comparable. Except for Kho et al.<sup>15</sup>, all groups used manual expert annotations rather than histopathology as

gold standard for their segmentation algorithms<sup>12–15</sup>. Instead of aiming to predict histopathology, these methods were optimised to match the performance of human experts, making their evaluation fundamentally different from ours. In contrast, histopathology provides a more clinically relevant ground truth of tumour extent, especially when using high-resolution PET images that may reveal tumour heterogeneity beyond the familiarity of nuclear medicine experts<sup>16–18</sup>.

Veluponnar et al.<sup>12</sup> and Yeung et al. also tested the performance of their DL models for margin assessment. Veluponnar et al. reached a sensitivity and a specificity of 96% and 76%, respectively, for predicting positive and close margins ( $\leq 2$  mm) for 86 patients. For the NST tumours we found a slightly lower sensitivity of 89%, with a higher specificity of 91%. However, the scores reported by Veluponnar et al.<sup>12</sup> might suffer from data leakage, as they seem to have optimised their ensemble of DL models on their test set. In addition, they only distinguish between wide ( $>2$  mm) and close ( $\leq 2$  mm) margins, which does not match the clinical “ink on tumour” definition. Yeung et al.<sup>13</sup> reached a sensitivity of 75% and a specificity of 43% when predicting the margin status for five breast specimens. While we reached a higher performance for the NST tumours, the data set used by Yeung et al.<sup>13</sup> is very small and might not provide reliable results. Unfortunately, F1 scores are rarely reported in literature in the context of margin assessment, while the data is often strongly imbalanced, such that metrics like sensitivity, specificity, and accuracy give an overly optimistic view.

Automatic margin assessment based on micro-PET-CT images of cancer specimens has been performed before. Moraitis et al.<sup>10</sup> proposed to use an iterative thresholding method to segment tumours in [<sup>18</sup>F]PSMA-1007 micro-PET-CT images of the prostate. They only evaluated this segmentation approach in terms of margin assessment performance. They reported a sensitivity of 83% and a specificity of 100% for the detection of positive margins for 18 prostate lesions in seven patients<sup>10</sup>. Despite the small data set, the results look promising. In future work, we could also investigate the performance of such iterative thresholding method in case of lumpectomy specimens.

This study contains several limitations related to the data. First, there might be slight inaccuracies in the data labels. For the segmentation labels, there might be imperfections in the tumour annotations created by the pathologist. Amgad et al.<sup>19</sup> reported a DSC of 0.87 when comparing annotations of triple negative BCa made by two senior pathologists. In addition, while in our previous work<sup>20</sup> we showed that accurate co-registration can be obtained between the micro-PET-CT images and WSIs, slight misalignment might still limit the accuracy of the segmentation labels. For the margin status labels, mistakes might occur due to sampling bias in the histopathological workup<sup>8</sup>. Second, the data sets used in this work are rather small. Especially for ILC and DCIS tumours, more data is required to enhance and validate ResU-Net performance. Due to the limited data sets, no meaningful conclusions on model performance could be drawn for the ILC and DCIS tumours in this work. Moreover, the application of the model is currently limited to segmenting NST, ILC, and DCIS tumours. It is assumed that the BCa type is known from preoperative biopsy, such that the most suited tumour segmentation model can be applied. In rare cases with mixed or unclear BCa types, a combination of the proposed models can still be used, but the performance might slightly decrease. The model might also incorrectly detect tumours in entirely benign specimens, as relative scaling might amplify noise in the micro-PET images in this case. Third, we only have 2D micro-PET-CT images with true tumour segmentations available, while the goal is to segment 3D tumours. Due to the micro-PET resolution of around 1 mm, the PET signal in a 2D image is the result of a 3D tissue configuration, which cannot be exploited by the 2D ResU-Net at this point. These 2D segmentation images also limit the accuracy of contour metrics like the 95HD and 1mmCD, as we cannot account for overlap or distances between contours across slices. Moreover, mismatches between the 2D lamella images used to train the ResU-Net, and the 3D images of entire BCa specimens used for model validation can also limit model performance. While we used the same imaging, reconstruction, and processing conditions for the fresh lamellas and entire BCa specimens to improve consistency

between our two data sets, minor differences remain. Some lamellas contain a clean cut edge made by the pathologist. While such straight cut edges are usually not found in entire BCa specimens, we believe the effect on model training and performance is negligible. In addition, the lamellas have a limited thickness of 2–3 mm, possibly leading to lower signals in the micro-PET images. However, we expect these partial volume effects to be minimal.

Before the proposed DL models could be integrated in the software of the micro-PET-CT specimen scanners and applied in clinical practice, some additional steps are required. Future research could focus on collecting additional data to further improve and validate the DL models. In particular for ILC and DCIS tumours, for which obtaining clear margins can be challenging in clinical practice, model performance should be further improved and validated. Collecting micro-PET-CT images with 3D true histopathological tumour segmentations may offer improved insight in the signature of ILC and DCIS tumours in micro-PET images. Methods like the ones proposed by Rusu et al.<sup>21,22</sup> could be explored to perform 3D co-registration between micro-PET-CT images and histopathology WSIs. Such 3D images with segmentations labels could be exploited to train and evaluate a 3D ResU-Net, possibly enhancing model performance. It may also be valuable to investigate whether the proposed models correctly predict the orientation of positive margins. In addition, it should be investigated how we can quantify the model’s confidence when making predictions. Finally, to further enhance clinical applicability, we might explore how the DL models can support the physicians to improve performance and efficiency for margin assessment, and we might consider faster, lightweight models for potential intraoperative use.

In conclusion, we trained and evaluated a ResU-Net to segment (pre) malignant breast tissue and predict the margin status based on micro-PET-CT images of lumpectomy specimens. For the segmentation of invasive carcinoma of the NST type, the ResU-Net showed a higher PR-AUC value compared to conventional intensity thresholding. In terms of margin status prediction for NST tumours, the ResU-Net matched the majority vote of seven physicians, and outperformed intensity thresholding. The fact that the fully automatic ResU-Net could match the performance of a panel of seven physicians, highlights its potential as a clinical support tool when performing IMA based on micro-PET-CT images of lumpectomy specimens, possibly improving consistency and reducing decision-making time. Although the results look promising, more extensive validation of the ResU-Net (in a clinical context) is required. This DL model is a first step towards a decision-support system to enhance micro-PET-CT-based IMA in BCa, facilitating the adoption of this technique in clinical practice.

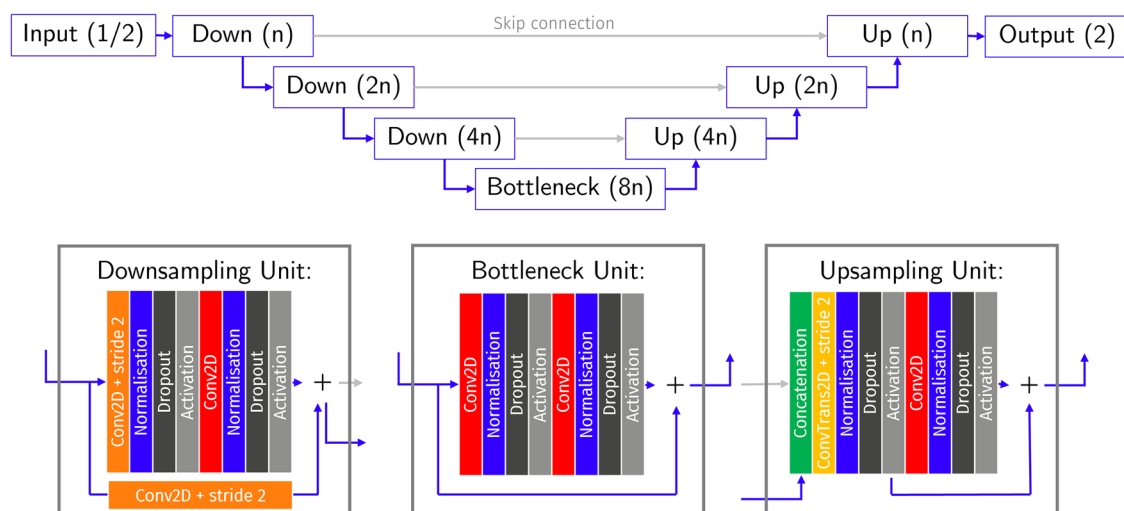
## Methods

### Data collection: study design

All data used in this manuscript were collected at the Ghent University Hospital in two clinical trials (ClinicalTrials.gov identifiers: NCT04343079 and NCT04999917), approved by the Ghent University Hospital ethics committee and conducted in accordance with the Declaration of Helsinki. We collected micro-PET-CT images and histopathological data from specimens of patients with early-stage BCa of the NST, ILC, or DCIS type undergoing BCS. All patients provided written informed consent for study participation and publication of the results.

### Micro-PET-CT specimen images

Patients who participated in the study received a preoperative injection with either 0.8 MBq/kg or 4 MBq/kg [<sup>18</sup>F]FDG. All micro-PET-CT images for this study were either acquired with the  $\beta$ -CUBE<sup>®</sup> and X-CUBE<sup>®</sup> (MOLE-CUBES NV, Belgium), or the AURA 10<sup>®</sup> (XEOS Medical NV, Belgium). While the  $\beta$ -CUBE<sup>®</sup> and X-CUBE<sup>®</sup> are meant for preclinical in vivo imaging, and the AURA 10<sup>®</sup> is dedicated to ex vivo specimen scanning, both scanners have similar specifications, and offer similar image quality with a sub-millimetre spatial resolution for PET. To obtain a uniform data set, all images were reconstructed to emulate a PET acquisition time of 10 min and an injected dose of 0.8 MBq/kg. PET images were reconstructed using 20 iterations of the maximum likelihood expectation maximisation (MLEM)



**Fig. 5 | Illustration of the Residual U-Net (ResU-Net) architecture used in this work.** The composition of the residual downsampling, bottleneck, and upsampling units is shown. The number of layers shown in this figure is only indicative, as a suited number of layers is defined using a grid search. The number of

channels ( $n$ ) is doubled for every downsampling unit and halved for every upsampling unit. We used the MONAI<sup>30</sup> implementation of the architecture introduced by Kerfoot et al.<sup>31</sup>.

algorithm, with an energy window of 50% around the 511 keV photopeak and an isotropic voxel size of 400  $\mu\text{m}$ . CT acquisition took 3 min with the X-CUBE<sup>®</sup> and 1 min with the AURA 10<sup>®</sup>. CT images were reconstructed using the image space reconstruction algorithm (ISRA) (X-CUBE<sup>®</sup>) or filtered back projection (FBP) (AURA 10<sup>®</sup>), with an isotropic voxel size of 100  $\mu\text{m}$ . The results were co-registered PET and CT images in DICOM format.

Several preprocessing steps were applied to the micro-PET-CT images before feeding them to the model. The 3D micro-CT images were expressed in Hounsfield units (HUs) and the 3D micro-PET images were converted to lean body mass standardised uptake values (SUVs) using the Janmahasatian formulation<sup>23</sup>. All micro-PET images were corrected for attenuation and median filtered with a 2 mm isotropic kernel to remove noise. DICOM images were converted to NUMPY arrays. The micro-PET images were interpolated to the grid of the micro-CT images with a third-order interpolation. The mouse beds, or specimen containers, were removed from the micro-CT images, after which the images were thresholded using Otsu’s method to define a binary tissue mask<sup>24</sup>. Only the largest connected tissue structure was selected, and the obtained tissue mask was used to remove the background from the micro-CT images and crop the micro-PET-CT images to only include the tissue.

**Ground truth image labels**

For this work, we collected two data sets with different types of ground truth labels derived from histopathology. The first data set consisted of 2D micro-PET-CT slices of lumpectomy specimens with the corresponding true semantic tumour annotations. This data set allowed supervised training and testing of the DL model for tumour segmentation. To obtain this data set, we collected micro-PET-CT images of lamellas of lumpectomy specimens. Immediately after resection, the fresh tumour specimens were cut into lamellas of around 2–3 mm thickness, and two to three lamellas were imaged. Afterwards, each lamella was further processed following clinical routine to obtain one hematoxylin and eosin (H&E) stained tissue section, of which a digital WSI was captured (NanoZoomer<sup>®</sup>, Hamamatsu Photonics, Japan). The collected micro-PET-CT images and the corresponding WSIs were co-registered using an in-house developed algorithm, as previously published<sup>20</sup>. Per lamella, this resulted in one 2D micro-PET-CT image, aligned with the 2D histopathology WSI.

A pathologist with more than 15 years of experience in breast pathology annotated the NST, ILC, and DCIS tumours in the WSIs using QuPath<sup>25</sup>. As such, we obtained binary tumour annotations, distinguishing

between (pre)malignant and benign breast tissue. For the NST cases we focused on the invasive tumour component, and did not include the DCIS components. Due to small co-registration inaccuracies, as well as tears and folds in the WSIs, the tissue in the micro-PET-CT images and WSIs did not always perfectly overlap. Hence, pixels that did not show tissue in the micro-CT image were automatically classified as benign. Pixels that showed tissue in the micro-CT image, but not in the WSI, were labelled as “unknown” and were ignored during model training and testing. Holes in the obtained tumour segmentations were filled as we are mainly interested in correctly predicting the tumour contours for margin assessment.

The second data set consisted of 3D micro-PET-CT images of lumpectomy specimens with known true histopathological margin status. This data set was used to validate the performance of the DL model in terms of margin assessment. To obtain this data set, we collected micro-PET-CT images of entire lumpectomy specimens immediately after they were resected from the patient. At the histopathology department, these specimens were processed following standard-of-care and the margin status was assessed following ASCO-CAP protocols<sup>26,27</sup>. This data set was previously described by De Crem et al.<sup>9</sup>. In accordance with De Crem et al.<sup>9</sup>, we used the same “ink on tumour” definition of a positive margin for all BCa types, and in case of NST tumours we only accounted for the invasive tumour component.

**Deep learning model**

A commonly used DL model architecture for semantic segmentation is U-Net and its variants<sup>28</sup>. In this work, we trained and evaluated a Residual U-Net (ResU-Net) to segment (pre)malignant breast tissue in micro-PET-CT lumpectomy images. ResU-Net was previously successfully applied in various biomedical image segmentation applications<sup>29</sup>. We used the MONAI<sup>30</sup> implementation of the ResU-Net, as introduced by Kerfoot et al.<sup>31</sup>. The building blocks of this model are residual down- and upsampling units, connected with each other through skip connections. Each block contains two convolutional layers, each followed by a normalisation, dropout, and activation layer. The usage of residual units helps to mitigate vanishing gradients, enabling training of deeper networks with faster convergence<sup>29,31</sup>. Figure 5 visualises the composition of the blocks. The model takes either a micro-CT, a micro-PET, or micro-PET-CT image as input and provides a probability map, indicating a pixel-wise tumour probability (between 0 and 1), as output. All code for this work was implemented using PyTorch<sup>32</sup>.

The first data set, containing the tumour segmentation labels, was used to train and test the ResU-Nets for tumour segmentation. This first data set

was split in train, test, and validation sets using a stratified group nested five-fold cross-validation scheme, with both outer and inner loops containing five folds. Before feeding the images to the model, they were preprocessed. First, the true tumour segmentations were one-hot encoded. Second, the micro-PET and micro-CT images were normalised to a range of [0,1] using min-max scaling. As HUs are considered to be quantitative units, the 5th and 95th percentiles of the combined foreground voxel intensities of all train images were used as the lower and upper limit to perform absolute scaling of the micro-CT images for all patients. As SUVs are considered to be semi-quantitative units, 0 and 80% of the maximal SUV per patient were used as the lower and upper limit for relative patient-specific scaling of the micro-PET images. Intensities outside the defined range were clipped to improve image contrast. In addition, we also implemented image augmentation. In every epoch, random patches were extracted from all train images. Selected image patches could undergo random rotations, flipping, or image degrading effects (addition of Gaussian noise, Gaussian smoothing, or contrast adjustment). When using the trained ResU-Nets to make predictions for test cases, the test images were min-max scaled and clipped using the same lower and upper limits as defined for the train set. Clipping is especially important to suppress high intensities in the micro-CT images when making predictions for specimens that contain metal wires. In addition, the test images were padded with background pixels up to the smallest size suited to pass through the network.

The ResU-Net was initialised using the standard Kaiming initialisation<sup>33</sup>. PReLU activations were implemented throughout the network, except for the last layer, where a Softmax activation was used<sup>33</sup>. We used the dice loss with background exclusion and batch-wise summation of intersection and union, as implemented in MONAI<sup>30</sup>. The stochastic gradient descent (SGD) optimisation algorithm with Nesterov momentum was used<sup>34</sup>, and a learning rate scheduler was implemented, reducing the learning rate with a factor of two every time a plateau was reached in the validation DSC. To tune other model hyperparameters, we performed a grid search using the Optuna framework<sup>35</sup>. Variations were made in the number of down- and upsampling layers in the network, the number of channels in the first layer, the convolutional kernel size, the type of normalisation layers (instance, batch, layer, or group normalisation), the dropout ratio, the learning rate, the batch size, the patch size, the number of random patches extracted per image every epoch, the bias towards selecting a patch with a positive pixel as centre coordinate, and the magnitude of the L2 penalty. Per test fold, we compared the results of 100 hyperparameter combinations, chosen through a Bayesian approach. For every hyperparameter combination, we trained five models (one per validation fold), each for 150 epochs. Per test fold, we chose the hyperparameter combination for which on average, across the five validation folds, the highest validation DSC was reached. The result were five hyperparameter combinations, one per test fold.

Using the hyperparameters as defined through a grid search, we trained five models per test fold. To avoid overfitting of the models, we used the validation data to perform early stopping. We trained models for three different inputs: (1) CT-only input, (2) PET-only input, (3) dual channel PET-CT input. Exactly the same data splits were used for all three model inputs, and separate grid searches were performed. All models were first trained using only the NST data samples. Afterwards, we performed transfer learning to fine-tune the models for ILC and DCIS segmentation. Model training was conducted on an NVIDIA V100 GPU with 32 GB VRAM.

### Intensity thresholding model

To gain more insight into the value of using a DL model in this task, we compared the performance of the ResU-Net to a conventional intensity thresholding approach. We used the same stratified group five-fold train-test split to find the most optimal intensity thresholds to segment the (pre) malignant breast tissue in micro-PET-CT images. The train set was used to define the threshold that on average resulted in the highest DSC when comparing the predicted and true tumour segmentations. For the micro-CT images, an absolute threshold expressed in HUs was defined. For every train set, we varied the threshold between  $-2000$  and  $+2000$  HUs with a step size

of 2, and chose the threshold that on average produced the highest DSC for these train samples. In case of micro-PET images, we used the common approach of defining a relative threshold per patient, expressed as a percentage of the maximum SUV<sup>16,17,36–38</sup>. For every train set, we varied the threshold between 0% and 100% of the maximum SUV per patient, and chose the relative threshold that on average produced the highest DSC for these train samples. For the combined micro-PET-CT images, we first identified high density tissues in the micro-CT images using Otsu's thresholding method<sup>24</sup>. The obtained high density regions were then used to mask the micro-PET signal, before defining the relative SUV threshold following the same approach as explained before for the micro-PET images. Separate thresholds were defined for the NST, and the combined ILC and DCIS tumours.

### Model performance evaluation: tumour segmentation

To evaluate the performance of the ResU-Nets and intensity thresholding models for tumour segmentation, we used the data set with true tumour segmentation labels. In case of the ResU-Net, we averaged the predictions of the appropriate ensemble of five trained models to obtain an average tumour probability map per test case. The final predicted tumour segmentation was obtained by thresholding the tumour probability map at 0.5, and performing hole-filling. In case of intensity thresholding, we obtained tumour probability maps by applying min-max scaling to the micro-CT or micro-PET images using the minimum and maximum intensity value per patient as the lower and upper limit, respectively. The final tumour segmentation was obtained by applying the appropriate optimal threshold, as found for the train set, to the input test image and performing hole-filling.

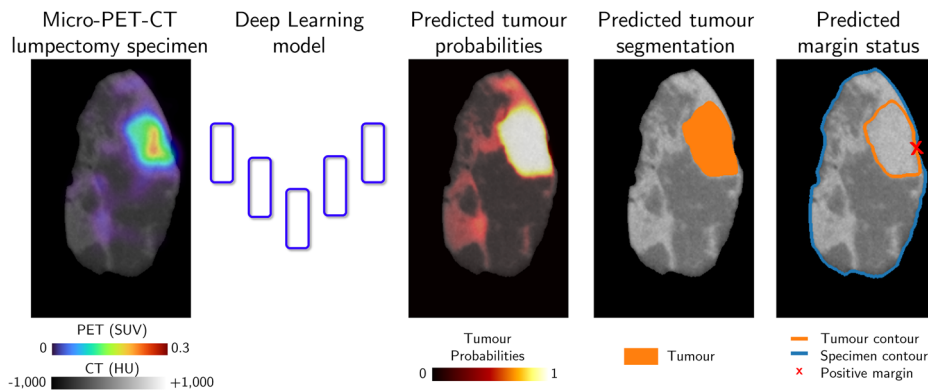
To compare the predicted tumour probability maps and tumour segmentations with the true histopathological tumour segmentations, we calculated the DSC and PR-AUC. The DSC ranges between 0 and 1 and scores the overlap between the true and predicted tumour segmentation. It is a well-suited metric to evaluate segmentation performance in medical images, which often show extensive class imbalance<sup>39</sup>. In addition, as we want to use the tumour segmentations to predict the margin status, it is important that the tumour contour is accurately predicted. To compare the predicted and true tumour contours, we calculated the 95HD, as well as the 1mmCD. For the 1mmCD we implemented a 2D variant of the surface dice as implemented in MONAI<sup>30</sup>. All metrics were calculated per patient, by combining the predictions for all 2D slices that belong to this patient. It should be noted that for the calculation of the 95HD, slices that only contained a predicted or true tumour segmentation could not be taken into account as the relative positions of the 2D image slices in the 3D volume is unknown.

To assess differences in tumour segmentation performance between the ResU-Nets and intensity thresholding models, the Wilcoxon signed-rank test was applied. Pairwise  $p$ -values were calculated for each evaluation metric, stratified per BCa type (NST, ILC, and DCIS), and per model input (micro-CT, micro-PET, and micro-PET-CT). A significance level  $\alpha = 0.05$  was used, and a Bonferroni-Holm correction was performed within each BCa type and model input, to compensate for multiple testing across the four evaluation metrics ( $m = 4$ ). In addition, the Wilcoxon signed-rank test was also conducted to compare the performance of the ResU-Nets and intensity thresholding models across different model inputs: micro-PET, micro-CT, and micro-PET-CT. Pairwise  $p$ -values were calculated for each evaluation metric and each comparison of model inputs, stratified per BCa type and model type (ResU-Nets and intensity thresholding models): (1) micro-PET vs. micro-CT, (2) micro-PET-CT vs. micro-CT, (3) micro-PET-CT vs. micro-PET. A significance level  $\alpha = 0.05$  was used, and a Bonferroni-Holm correction was performed within each BCa type and model type to compensate for multiple testing across the three model input comparisons and four evaluation metrics ( $m = 3 \cdot 4 = 12$ ).

### Model performance evaluation: margin assessment

To evaluate the performance of the ResU-Nets and intensity thresholding models in terms of margin assessment, we used the data set with known true

**Fig. 6 | Illustration of the workflow used to automatically segment the tumour and predict the margin status for a lumpectomy specimen with the deep learning (DL) model.** The DL model takes a 3D micro-PET-CT image as input, and predicts voxel-wise tumour probabilities. The predicted tumour probabilities (between 0 and 1) are here represented by a heatmap on top of the micro-CT. The predicted tumour segmentation, obtained by thresholding the predicted tumour probabilities at 0.5, is shown in orange on top of the micro-CT. A positive margin is found at the location where the tumour contour (in orange) touches the specimen contour (in blue). Every figure panel only shows one slice of the 3D image of the lumpectomy specimen.



histopathological margin status labels. Figure 6 visualises the workflow when using the DL model to predict the margin status for a micro-PET-CT lumpectomy image. In case of the ResU-Net, we used the appropriate ensemble of 25 trained models to predict tumour probability maps based on the 3D micro-PET-CT images. Tumour probability maps were predicted slice-per-slice in three orthogonal directions. All predicted probability maps were averaged per case across the three directions and across the 25 models, resulting in one predicted probability map per case. The final tumour segmentation was obtained by thresholding this probability map at 0.5. In case of intensity thresholding, we used the appropriate combination of five resulting intensity thresholds to predict tumour segmentations. These five predicted tumour segmentations were averaged per case and thresholded at 0.5 to obtain the final tumour segmentation.

To determine the predicted margin status, we defined the minimum distance between the outer contour of the predicted tumour segmentation, and the specimen contour. The specimens were segmented by applying Otsu's thresholding to the micro-CT images as explained before<sup>24</sup>. When the minimum distance between the outer contour of the predicted tumour segmentation and the specimen contour was  $\leq 0.1$  mm, we considered the predicted margin status to be positive. This was done to take into account the uncertainty on the exact location of the specimen contour due to the limited spatial resolution of the micro-CT images. We compared these predicted margin statuses with the true histopathological margins statuses for all cases, and calculated the sensitivity, specificity, precision, and F1 score for the detection of positive margins. We compared the performance of the ResU-Net and intensity thresholding with the performance of seven physicians (three breast surgeons, one gynaecology resident, two nuclearists, and one radiologist) who interpreted the same set of micro-PET-CT images and predicted the margin status. Comparing the predicted and true histopathological margin status allowed to calculate the same evaluation scores for the physicians. We report both the evaluation scores obtained when using the majority vote of the seven physicians to define the margin status per breast specimen, as well as the average of the evaluation scores reached across the physicians. Details of this interpretation study are described by De Crem et al.<sup>9</sup>.

To evaluate differences in margin assessment performance between the ResU-Nets, intensity thresholding models, and the majority vote of seven physicians, McNemar's test was applied. Given the limited sample size, the exact version of McNemar's test was used. Pairwise  $p$ -values were calculated for each comparison of margin assessment strategies, stratified per BCa type: (1) Majority vote of seven physicians vs. ResU-Net prediction, (2) Majority vote of seven physicians vs. intensity thresholding prediction, (3) ResU-Net prediction vs. intensity thresholding prediction. A significance level  $\alpha = 0.05$  was used, and a Bonferroni-Holm correction was performed within each BCa type to compensate for multiple testing across the three comparisons in margin assessment strategies ( $m = 3$ ).

## Data availability

The data sets generated and analysed in the current study are not publicly available, but are available from the corresponding author on reasonable request and after approval of the ethics committee.

## Code availability

The underlying code for this study and trained deep learning models are not publicly available for proprietary reasons.

Received: 10 April 2025; Accepted: 13 July 2025;

Published online: 09 August 2025

## References

- Koopmansch, C. et al. Intraoperative evaluation of resection margins in breast-conserving surgery for in situ and invasive breast carcinoma. *Breast Cancer Res.* **15**, 117822342199345 (2021).
- Cardoso, F. et al. Early breast cancer: Esmo clinical practice guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* **30**, 1194–1220 (2019).
- McDonald, E. S., Clark, A. S., Tchou, J., Zhang, P. & Freedman, G. M. Clinical diagnosis and management of breast cancer. *J. Nucl. Med.* **57**, 9S–16S (2016).
- Moran, M. S. et al. Society of Surgical Oncology-American Society for Radiation Oncology consensus guideline on margins for breast-conserving surgery with whole-breast irradiation in stages I and II invasive breast cancer. *Int. J. Radiat. Oncol. Biol. Phys.* **88**, 553–564 (2014).
- Brouwer de Koning, S. G., Vrancken Peeters, M.-J. T. F. D., Józwiak, K., Bhairosing, P. A. & Ruers, T. J. M. Tumor resection margin definitions in breast-conserving surgery: systematic review and meta-analysis of the current literature. *Clin. Breast Cancer* **18**, e595–e600 (2018).
- Vos, E. L., Jager, A., Verhoef, C., Voogd, A. C. & Koppert, L. B. Overall survival in patients with a re-excision following breast conserving surgery compared to those without in a large population-based cohort. *Eur. J. Cancer* **51**, 282–291 (2015).
- Göker, M. et al. 18f-fdg micro-pet/ct for intra-operative margin assessment during breast-conserving surgery. *Acta Chir. Belg.* **120**, 366–374 (2020).
- Lambert, B. et al. Feasibility study on the implementation of a mobile high-resolution pet/ct scanner for surgical specimens: exploring clinical applications and practical considerations. *Eur. J. Nucl. Med. Mol. Imaging* **52**, 2979–2994 (2025).
- De Crem, A.-S. et al. Bimap: breast cancer intra-operative margin assessment using pet-ct. *ESMO Open* **9**, P89 (2024).
- Moraitis, A. et al. Evaluation of surgical margins with intraoperative psmat pet/ct and their prognostic value in radical prostatectomy. *J. Nucl. Med.* **66**, 352–358 (2025).

11. Masannat, Y. et al. Multifocal and multicentric breast cancer, is it time to think again? *Ann. R. Coll. Surg. Engl.* **102**, 62–66 (2020).
12. Veluponnar, D. et al. Toward intraoperative margin assessment using a deep learning-based approach for automatic tumor segmentation in breast lumpectomy ultrasound images. *Cancers* **15**, 1652 (2023).
13. Yeung, C. et al. From quantitative metrics to clinical success: assessing the utility of deep learning for tumor segmentation in breast surgery. *Int. J. Comput. Assist. Radiol. Surg.* **19**, 1193–1201 (2024).
14. Shia, W.-C. et al. Evaluating the margins of breast cancer tumors by using digital breast tomosynthesis with deep learning: a preliminary assessment. *Diagnostics* **14**, 1032 (2024).
15. Kho, E. et al. Broadband hyperspectral imaging for breast tumor detection using spectral and spatial information. *Biomed. Opt. Express* **10**, 4496 (2019).
16. Zaidi, H. & El Naqa, I. Pet-guided delineation of radiation therapy treatment volumes: a survey of image segmentation techniques. *Eur. J. Nucl. Med. Mol. Imaging* **37**, 2165–2187 (2010).
17. Hatt, M. et al. Classification and evaluation strategies of auto-segmentation approaches for pet: report of AAPM Task Group No. 211. *Med. Phys.* **44**, e1–e42 (2017).
18. Debacker, J. M. et al. Direct co-registration of [18f]fdg uptake and histopathology in surgically excised malignancies of the head and neck: a feasibility study. *Eur. J. Nucl. Med. Mol. Imaging* **50**, 2127–2139 (2023).
19. Amgad, M. et al. Structured crowdsourcing enables convolutional segmentation of histology images. *Bioinformatics* **35**, 3461–3467 (2019).
20. Maris, L. et al. Method for co-registration of high-resolution specimen pet-ct with histopathology to improve insight into radiotracer distributions. *EJNMMI Phys.* **11**, 85 (2024).
21. Rusu, M. et al. Co-registration of pre-operative ct with ex vivo surgically excised ground glass nodules to define spatial extent of invasive adenocarcinoma on in vivo imaging: a proof-of-concept study. *Eur. Radiol.* **27**, 4209–4217 (2017).
22. Rusu, M. et al. Registration of presurgical mri and histopathology images from radical prostatectomy via rapsodi. *Med. Phys.* **47**, 4177–4188 (2020).
23. Janmahasatian, S. et al. Quantification of lean bodyweight. *Clin. Pharmacokinet.* **44**, 1051–1065 (2005).
24. Otsu, N. A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern.* **9**, 62–66 (1979).
25. Bankhead, P. et al. Qupath: open source software for digital pathology image analysis. *Sci. Rep.* **7**, 16878 (2017).
26. Lester, S. C. et al. Protocol for the examination of specimens from patients with invasive carcinoma of the breast. *Arch. Pathol. Lab. Med.* **133**, 1515–1538 (2009).
27. Lester, S. C. et al. Protocol for the examination of specimens from patients with ductal carcinoma in situ of the breast. *Arch. Pathol. Lab. Med.* **133**, 15–25 (2009).
28. Ronneberger, O., Fischer, P. & Brox, T. U-net: convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*, Lecture Notes in Computer Science, 234–241 (Springer Nature, 2015).
29. Siddique, N., Paheding, S., Elkin, C. P. & Devabhaktuni, V. U-net and its variants for medical image segmentation: a review of theory and applications. *IEEE Access* **9**, 82031–82057 (2021).
30. Cardoso, M. J. et al. Monai: an open-source framework for deep learning in healthcare. (2022).
31. Kerfoot, E. et al. Left-ventricle quantification using residual u-net. In *Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges*, Lecture Notes in Computer Science, 371–380 (Springer Nature, 2019).
32. Paszke, A. et al. Pytorch: an imperative style, high-performance deep learning library. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 8026–8037 (Adv. Neural Inf. Process. Syst., 2019).
33. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778 (IEEE, 2016).
34. Nesterov, Y. A method for solving the convex programming problem with convergence rate  $O(1/k^2)$ . *Proc. USSR Acad. Sci.* **269**, 543–547 (1983).
35. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: a next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2623–2631 (ACM, 2019).
36. Mercieca, S. et al. Comparison of suvmax and suvpeak based segmentation to determine primary lung tumour volume on fdg pet-ct correlated with pathology data. *Radiother. Oncol.* **129**, 227–233 (2018).
37. Kim, M. et al. Evaluation of thresholding methods for the quantification of [68ga]ga-psma-11 pet molecular tumor volume and their effect on survival prediction in patients with advanced prostate cancer undergoing [177lu]lu-psma-617 radioligand therapy. *Eur. J. Nucl. Med. Mol. Imaging* **50**, 2196–2209 (2023).
38. Trada, Y. et al. Impact of tumour region of interest delineation method for mid-treatment FDG-PET response prediction in head and neck squamous cell carcinoma undergoing radiotherapy. *Quant. Imaging Med. Surg.* **13**, 2822–2836 (2023).
39. Müller, D., Soto-Rey, I. & Kramer, F. Towards a guideline for evaluation metrics in medical image segmentation. *BMC Res. Notes* **15**, 210 (2022).

## Acknowledgements

This project has received funding from VLAIO through a Baekeland mandate with project ID HBC.2022.0175. The funder played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

## Author contributions

L.M., M.G., K.D.M. and B.V.d.B. acquired the imaging data for this manuscript. M.G. performed the breast-conserving surgeries, and K.D.M. and B.V.d.B. performed the preoperative radiotracer injections. K.V.d.V. annotated all histopathology WSIs. M.G., K.D.M., and B.V.d.B. participated in the interpretation study where they predicted the margin status of lumpectomy specimens based on their micro-PET-CT images. L.M. developed and evaluated the deep learning models, and interpreted the results, with the support of V.K., C.V. and S.V.H. L.M. wrote the manuscript, and all authors commented on the manuscript. All authors read and approved the final manuscript.

## Competing interests

L.M. is a research engineer at XEOS Medical, and V.K. is a shareholder and board member of XEOS Medical. M.G., K.D.M., B.V.d.B., S.V.H., K.V.d.V., and C.V. declare that they have no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41523-025-00797-w>.

**Correspondence** and requests for materials should be addressed to Luna Maris.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025