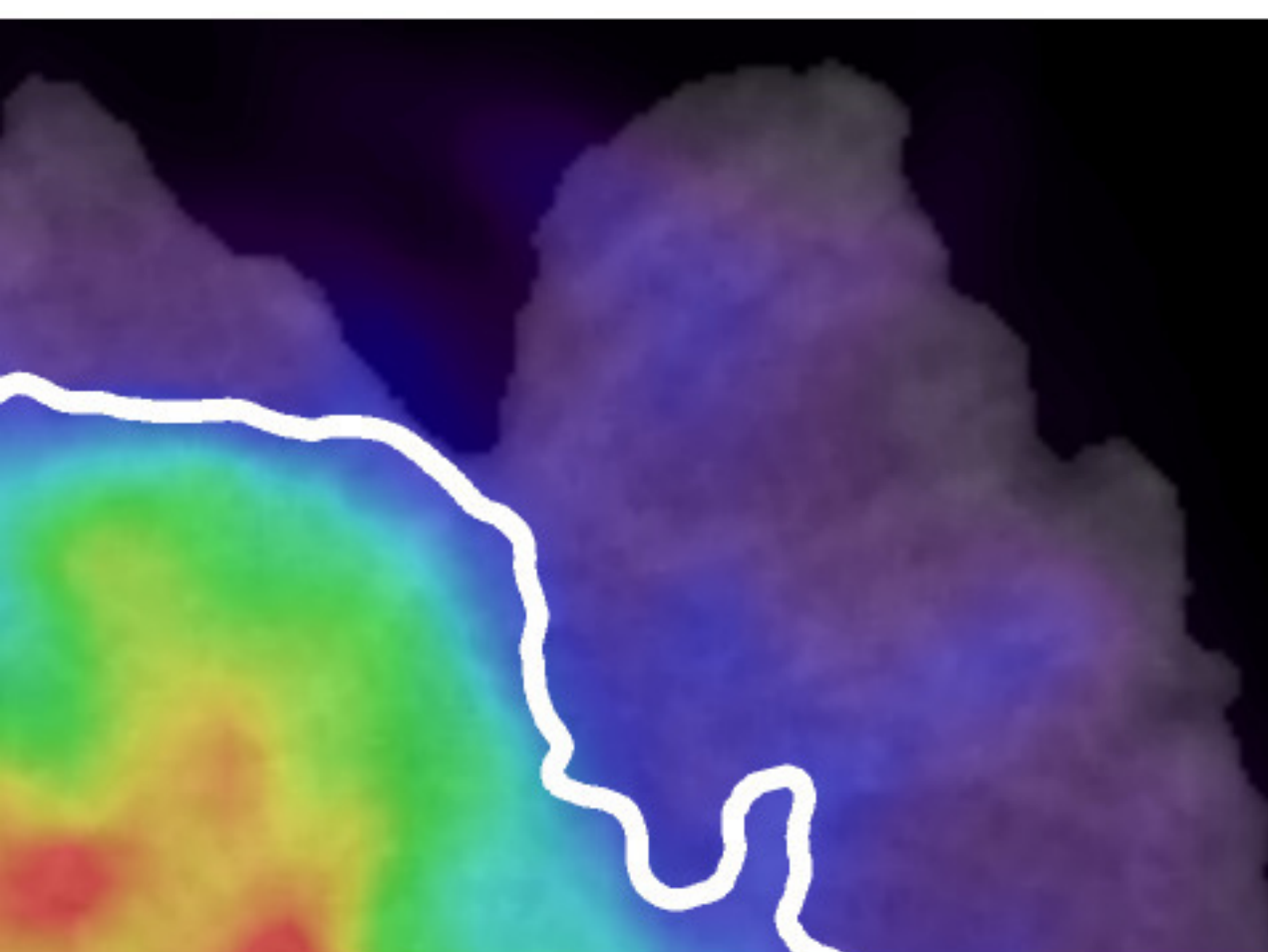




Deep Learning Analysis of Micro-PET-CT Images to Support Intraoperative Margin Assessment in Breast-Conserving Surgery

Luna Maris

Doctoral dissertation submitted to obtain the academic degree of
Doctor of Biomedical Engineering

Supervisors

Prof. Christian Vanhove, PhD* - Prof. Vincent Keereman, PhD* - Prof. Koen Van de Vijver, PhD**

* Department of Electronics and Information Systems
Faculty of Engineering and Architecture, Ghent University

** Department of Diagnostic Sciences
Faculty of Medicine and Health Sciences, Ghent University

September 2026

Deep Learning Analysis of Micro-PET-CT Images to Support Intraoperative Margin Assessment in Breast-Conserving Surgery

Luna Maris

Doctoral dissertation submitted to obtain the academic degree of
Doctor of Biomedical Engineering

Supervisors

Prof. Christian Vanhove, PhD* - Prof. Vincent Keereman, PhD* - Prof. Koen Van de Vijver,
PhD**

- * Department of Electronics and Information Systems
Faculty of Engineering and Architecture, Ghent University
- ** Department of Diagnostic Sciences
Faculty of Medicine and Health Sciences, Ghent University

September 2026

ISBN 978-94-93513-97-6

NUR 954, 984

Wettelijk depot: D/2026/10.500/105

Members of the Examination Board

Chair

Prof. Joris Degroote, PhD, Ghent University

Other members entitled to vote

Prof. Markus Hahn, PhD, Eberhard Karls Universität Tübingen, Germany

Bieke Lambert, PhD, AZ Maria Middelaers & AZ Jan Palfijn Gent

Prof. Sofie Van Hoecke, PhD, Ghent University

Prof. Jef Vandemeulebroucke, PhD, Vrije Universiteit Brussel

Prof. Sofie Verbeke, PhD, Ghent University

Supervisors

Prof. Christian Vanhove, PhD, Ghent University

Prof. Vincent Keereman, PhD, Ghent University

Prof. Koen Van de Vijver, PhD, Ghent University

Acknowledgements

Reaching the end of this PhD, I find myself looking back on a journey that was exciting, sometimes challenging, and above all rewarding. A PhD is not only about the research itself, it is also an opportunity to grow as a person. I learned how to manage a project, how to cope with the ups and downs, and, most of all, to persevere. I could not have completed this work on my own, and I would like to thank everyone who helped me get here, through their guidance, their support, or simply by being there when I needed them most. To all of you: a sincere thank you.

This journey began five years ago, during my master's thesis in biomedical engineering, where I discovered my passion for machine learning in medical image analysis. When I told my master's thesis promoter, dr. Roel Van Holen, that I wanted to pursue a PhD in this field, he did not hesitate to offer me a position as Research Engineer at his start-up company, XEOS, where this PhD took shape. **Roel**, thank you for your confidence in me when you offered me this incredible opportunity. That trust set everything in motion.

During my PhD, I could always rely on the guidance of my promoters: Prof. Vincent Keereman, Prof. Christian Vanhove, and Prof. Koen Van de Vijver. Whenever I got lost, they encouraged me to take a step back and look at things from a different perspective. Their enthusiasm was contagious, and they continuously inspired me to engage with new research topics. **Vincent**, thank you for your unwavering support during our follow-up meetings, and for always fitting me into your incredibly busy schedule. You were always happy to think things through with me whenever I dove into something new, and you could always offer a fresh perspective and invaluable feedback. **Chris**, thank you for showing me the way in the lab and for guiding me through my first acquisition of a micro-PET-CT image of a breast specimen. You were always calm and down-to-earth, and I could count on you for feedback at any time. You are also, undoubtedly, the fastest manuscript reviewer of all time. **Koen**, we spent countless hours together reviewing histopathology slides. Thank you for your patience, for always making time to answer my questions, and for showing me around the histopathology lab. You speak about your work with such passion that it was a genuine pleasure to learn from you. I have watched you cut up your entire lunch on more than one occasion, just to demonstrate how a breast specimen is sliced. I truly admire your dedication.

I would also like to thank the members of my doctoral guidance committee, Prof. Koen Van Laere and Prof. Sofie Van Hoecke, for their constructive feedback on my research and for helping me stay on track. A special word of thanks goes to Prof. Van Hoecke and her research group, who gave me valuable input on the machine learning side of my PhD. They were always

deeply engaged during our meetings and offered ideas that helped shape new research directions. My gratitude also goes to the chair of my jury, Prof. Joris Degroote, and to the jury members, Prof. Markus Hahn, dr. Bieke Lambert, Prof. Sofie Van Hoecke, Prof. Jef Vandemeulebroucke, and Prof. Sofie Verbeke. They read my manuscript carefully and provided detailed comments, which helped me improve the quality and clarity of my final dissertation. Our discussions were truly stimulating and further broadened my view on the field.

To my co-authors, dr. Menekse Göker and Prof. Jens Debacker, thank you for helping me conduct my research and shape my publications. **Menekse**, thank you for your ever-present smile, no matter how busy your schedule was. You always made me feel welcome, you helped me find my way in the operating room, and taught me all about breast surgery. **Jens**, I started my very first research project with you. Your drive to do research and your determination to solve the most challenging problems are truly inspiring. Beyond my co-authors, I am grateful to all the breast surgeons, nuclear medicine physicians, and residents at the histopathology lab of Ghent University Hospital, who made it possible for me to collect the data this project relied on. I would also like to thank the MEDISIP research group at Ghent University, whose meetings were a constant source of inspiration. **Melissa** and **Florence**, I believe our first real conversation took place while we were wiping the sticky floors of the IbiTech room after our starters' party, I guess nothing quite creates a bond like that. Thank you both for always being willing to help and to answer any questions I had.

Throughout this PhD, the XEOS office was my second home, where I was surrounded by colleagues who were unfailingly supportive and genuinely interested in my research. **Leen** and **Anneleen**, I began this journey alongside you in the clinical team. You supported me in every way imaginable, from helping me collect data for my research to sharing your own PhD experiences, not to mention the occasional life lesson written on a post-it note and, of course, the memorable Easter egg tastings. Later, I made the switch to the R&D team. **Gert** and **Jakob**, you carried me through the final stages of my PhD, showed real interest in my work, were always ready to brainstorm with me, and helped me shape and improve my final manuscript. To all my other XEOS colleagues: together you form a dedicated, warm, and remarkably bright group that I am honoured to be part of. During my time as a PhD student, I also had the opportunity to guide many thesis students and interns at XEOS. It was always a pleasure to share my experience while learning just as much from them. Thank you, Roel, Anastasiia, Lore, Robine, Elly, Vita, Emile, Laure, and Vic for your enthusiasm, curiosity, and eagerness to learn.

A PhD does not always run smoothly, and during the inevitable difficult periods, you need friends and family to keep you going. **Emma** and **Kate**, you have been there from the very beginning, and I always looked forward to our dinner dates. As fellow PhD students, no one understood better what

I was going through. Thank you for listening, for the occasional shared cry, but most of all for laughing with me, because there was rarely a setback that a good cup of tea and a recital of the IKEA catalogue from memory could not solve. **Mariken** and **Nathalie**, we have known each other since kindergarten, and even though we don't see each other as often anymore, it always feels familiar and comforting the moment we do. You understand me so well, and I hope that never changes. **Louise**, our weekly therapeutic walking sessions simply cannot be left off this list. Thank you for sharing your advice, for cheering me up, and most of all for just listening.

I would like to thank my family for their unwavering support. Mom and Dad, **Ine & Johan**, thank you for always supporting me and for encouraging me to find a field of study that truly suited my interests. Even when that meant standing by me through more than a few tears and doubts as I decided to leave physics behind and pursue a master's in biomedical engineering. Thank you for being patient with me and for believing in me. **Fiebe**, my sister, thank you for sharing my, sometimes particularly weird, sense of humour, for making me laugh, and for reminding me that there is more to life than work. **Fons**, you are the fluffiest living rubber duck I could ever have imagined. To my grandparents, **Germana & Rik**, thank you for being my greatest supporters, and for always being there when I needed you. To **Muriel & Luc**, **Chris**, and **Chiara & Robbe**, thank you for welcoming me so warmly, and for the endless lunches and dinners, chamomile teas, and chocolates along the way.

Finally, I want to save my deepest gratitude for my boyfriend, who has been by my side for the entire journey. **Steven**, you have shared these years with me, both the highs and the lows, the excitement and the struggles, all of it. You were always there to offer advice, or simply to listen while I explained, at length, the latest ideas and solutions I had dreamed up for problems you never knew existed. Whenever I said I only needed five more minutes, you just quietly accepted that those five minutes could easily turn into an hour or two when I had unexpectedly disappeared down another rabbit hole. Even through the most stressful periods, when I was not always the easiest person to be around, you just sat beside me, helped me, and encouraged me. There are no words to express how much I appreciate your endless love, your patience, and the way you always manage to cheer me up. Above all, thank you for believing in me, even in the moments when I did not believe in myself.

Looking back, I feel nothing but gratitude, and I could not have wished for better people with whom to share this journey.

Luna Maris
Ghent, August 2026

Summary

Breast-conserving surgery (BCS), followed by radiotherapy, has become the standard treatment for most early-stage breast cancer (BCa) patients. Successful BCS requires complete tumour excision with a margin of surrounding healthy tissue, i.e. obtaining negative resection margins. Positive margins, indicating residual tumour cells at the specimen border, remain the main risk factor for local recurrence after BCS. Sectional histopathology is the gold standard for margin assessment. According to the 2014 Society of Surgical Oncology and American Society for Radiation Oncology (SSO-ASTRO) consensus, a tumour-free margin greater than 0 mm is considered adequate for invasive cancer of no special type (NST) and invasive lobular carcinoma (ILC), while a minimum margin of 2 mm is recommended for ductal carcinoma in situ (DCIS). Positive margins are found in approximately 20% of patients after initial BCS, with wide variations reported in literature. Re-excision is often required for these patients to clear the breast of remaining tumour cells, a procedure associated with psychological, physical, and economic consequences. To avoid secondary surgery, techniques for fast intraoperative margin assessment (IMA) are required. However, existing methods like frozen sections, imprint cytology, intraoperative ultrasound, and specimen radiography have important limitations. Currently, there is no cost-effective IMA method available that shows sufficient accuracy and does not disrupt the surgical workflow.

Recently, a promising technique for IMA has emerged: *ex vivo* specimen imaging with high-resolution (micro-) positron emission tomography (PET), combined with computed tomography (CT). A compact and mobile micro-PET-CT specimen scanner can generate 3D images of excised BCa specimens directly in the operating theatre, providing detailed information on both margin status and margin width. Clinical studies demonstrated the feasibility and safety of this technique, and showed that micro-PET-CT imaging reached a significantly higher success rate in addressing positive margins in BCa specimens compared to standard-of-care IMA techniques. In these studies, physicians manually interpreted the images in approximately one minute per 3D image. While combining assessments from multiple surgeons improved reproducibility, variability in single-reader evaluations remained. To further improve interpretation speed, inter-reader reliability, and possibly diagnostic performance, this dissertation explores first steps towards a decision-support system to assist physicians in interpreting micro-PET-CT images. For this purpose, a deep learning (DL) model was developed that automatically segments breast tumours in micro-PET-CT images of lumpectomy specimens, thereby also enabling automated margin status prediction.

To develop such a DL model, first, we collected a dataset of micro-PET-CT images of BCa specimens with the corresponding ground-truth tumour segmentations. For this purpose, we developed a novel method to accurately and automatically align micro-PET-CT specimen images with the corresponding histopathology whole slide images (WSIs). The proposed co-registration method combines an image acquisition scheme that minimises and tracks tissue deformation with a mathematical co-registration algorithm that further optimises image alignment. The image acquisition scheme consists of three steps: (1) Micro-PET-CT images of approximately 2 mm thick lamellas of fresh surgical specimens, enclosed in tissue cassettes, are acquired. (2) Micro-CT images of the same lamellas after formalin fixation are captured to visualise tissue deformation. (3) One haematoxylin and eosin (H&E) stained histopathology section per lamella is obtained of which a digital WSI is acquired. Next, an automatic co-registration algorithm was developed to improve the alignment between the micro-PET-CT images of the fresh lamellas and WSIs, guided by the micro-CT of the fixated lamellas. Compared with existing *in vivo* PET-histopathology alignment methods, the proposed workflow is less labour-intensive and time-consuming, requiring only fresh cutting of lamellas in addition to standard histopathology, and performing fully automatic co-registration using intensity-based similarity metrics.

We applied the co-registration method to 53, 19, and 10 cassettes from 19 NST, 6 ILC, and 3 DCIS specimens, respectively, from patients pre-operatively injected with 0.8 MBq/kg [^{18}F]fluorodeoxyglucose ([^{18}F]FDG). While the cassette framework minimised tissue deformation, the best alignment between the micro-PET-CT images and WSIs was achieved after deformable co-registration, indicating the need to correct for free-form tissue deformations. Using microcalcifications as landmarks, the final mean spatial co-registration error between micro-PET and histopathology was 0.69 ± 0.23 mm, which is comparable to the spatial resolution of the micro-PET images. The obtained 2D co-registered images allowed us to study standardised uptake values (SUVs) of different breast tissues, as identified on the WSIs by an experienced breast pathologist. In general, (pre)malignant tissue (including NST, ILC, and DCIS) showed higher SUVs than healthy tissue (including healthy glandular, connective, and adipose tissue). This suggests that ILC and DCIS tumours can be visualised using micro-PET-CT, while they are in general more difficult to detect in *in vivo* PET. As expected, inflamed tissue and skin also showed high uptake. Furthermore, tumour annotations from histopathology could be transferred to the aligned 2D micro-PET-CT images to obtain ground-truth breast tumour segmentations, crucial for the development and validation of DL-based segmentation models. While currently we only applied this method to (highly deformable) BCa specimens, we believe the proposed method is applicable to a wide range of specimens and radiotracers, providing insight into distributions of (new) radiotracers in human malignancies at a submillimetre resolution.

Afterwards, we used the collected dataset of 2D micro-PET-CT images with histopathology-derived ground-truth segmentations to develop a DL model for automated breast tumour segmentation. We used a stratified group nested five-fold cross-validation scheme to train and evaluate an ensemble of 2D Residual U-Nets (ResU-Nets) to delineate (pre)malignant tissue in micro-PET-CT images of breast cancer specimens. The models were first trained on the largest group of NST cases, after which transfer learning was performed for the combined group of ILC and DCIS cases. For the segmentation of NST tumours, the ResU-Nets reached a Dice similarity coefficient (DSC) of 0.73 ± 0.20 , and the area under the curve of the precision-recall curve (PR-AUC) was 0.83 ± 0.14 . We compared the performance of the ResU-Nets with conventional PET intensity thresholding, and found that, for the NST cases, the ResU-Nets outperformed intensity thresholding for the PR-AUC value. The segmentation of ILC and DCIS tumours was found to be more challenging, with DSC values of 0.49 ± 0.26 and 0.03 ± 0.02 , and PR-AUC values of 0.60 ± 0.25 and 0.03 ± 0.03 for ILC and DCIS, respectively. Although the ResU-Nets showed better performance compared to intensity thresholding for the ILC and DCIS cases, due to the small datasets, no statistically significant improvements were found. Furthermore, we showed that combining micro-PET and micro-CT information is essential for accurate tumour delineation, with dual-modality input models outperforming single-modality input models. This indicates that the combination of metabolic information from micro-PET, with high-resolution structural information from micro-CT, is crucial for accurate tumour segmentation. A unique aspect of our study is the use of histopathology as the gold standard for tumour segmentation instead of the more common manual expert annotations, providing a more clinically relevant ground truth of tumour extent.

Next, we validated the ability of the ResU-Nets to accurately predict the margin status of BCa specimens. For this purpose, we used the datasets collected in the BIMAP and BrIMA studies, which contain 3D micro-PET-CT images with histopathological ground-truth margin status. The BIMAP study collected images of 41 specimens at the Ghent University Hospital, whereas the BrIMA study imaged 152 specimens across six hospitals in three European countries. Both studies imaged early-stage BCa specimens (including NST, ILC, DCIS, and neoadjuvant-treated (NAT) specimens), with NST being the predominant subtype with 31 and 84 cases in BIMAP and BrIMA, respectively. The ensemble of ResU-Nets was used to segment (pre)malignant tissue in these previously unseen images, after which the margin status was predicted based on whether the segmented tumour touched the specimen border. When comparing the predicted and true histopathological margin status in all NST cases, F1 scores of 84% and 41% were reached for the BIMAP and BrIMA datasets, respectively. For all BCa subtypes combined, F1 scores of 76% and 50% were obtained for the BIMAP and BrIMA datasets. The performance of the

ResU-Nets was compared with conventional PET intensity thresholding, as well as with the intraoperative margin assessments of physicians interpreting the same micro-PET-CT images in the BrIMA study, and with the postoperative majority vote of seven and four physicians in the BIMAP and BrIMA studies, respectively. Both for the NST tumours and for the combined group of all BCa subtypes in the BIMAP and BrIMA datasets, the ResU-Nets showed margin assessment performance statistically equivalent to intraoperative and postoperative physicians' assessments, while consistently outperforming intensity thresholding. Notably, the ResU-Nets achieved physician-level performance on a large multicentre dataset containing diverse BCa subtypes, representative of the true clinical population. Although statistical significance could not always be shown for smaller ILC, DCIS, and NAT subsets, the ResU-Nets generally showed higher specificity, precision, and F1 scores compared to intensity thresholding. Moreover, despite not being trained on NAT tumours, the models also achieved performance comparable to physicians for this subgroup, indicating their robustness.

Finally, although the developed DL models showed promising performance, safe clinical deployment requires insight into prediction uncertainty, which is not captured by binary tumour segmentations or margin status labels. The tumour probability maps predicted by the DL models express this uncertainty at the pixel level. To make these probabilities reliable, we calibrated them, improving the trustworthiness and interpretability of the model outputs. Tumour probability maps predicted by the ResU-Nets were calibrated using the inductive cross-Venn-ABERS method, focusing on the largest subgroup of patients with NST BCa. Calibration was performed and evaluated on the dataset of 53 2D micro-PET-CT images from lumpectomy lamellas of 19 patients with NST BCa, using histopathology-derived tumour segmentations as ground truth. The validation folds in the group nested five-fold cross-validation scheme, used to train the segmentation models, were used as calibration sets. Calibration substantially improved agreement between predicted pixel-level tumour probabilities and observed frequencies of malignant pixels. The expected calibration error (ECE), binary cross-entropy (BCE) loss, and Brier loss were reduced by 60.5%, 17.7%, and 14.8%, to final values of 0.017, 0.172, and 0.052, respectively. Tumour segmentation performance, in terms of area under the curve of the receiver operating characteristic (ROC-AUC) and PR-AUC, remained stable after calibration. Calibrated probability maps were also generated for the 84 3D micro-PET-CT images of NST specimens in the BrIMA dataset, with true histopathological margin status labels. In terms of margin assessment performance, calibration led to a slight decrease in PR-AUC. Qualitatively, the probability maps provided transparent, spatially-resolved visualisations of prediction uncertainty, and calibration made these estimates more reliable, enabling trustworthy identification of uncertain tumour boundaries and ambiguous margin re-

gions. These findings suggest that calibrated tumour probability maps can support more informed, efficient, and risk-aware clinical decision-making.

In conclusion, the proposed DL models are capable of predicting both tumour segmentations and margin status labels, as well as communicating spatially-resolved prediction uncertainty. These models show great potential to improve consistency and reduce decision-making time when interpreting micro-PET-CT images of lumpectomy specimens. Although the results are promising, further validation is required before the DL models can be adopted in clinical practice. Larger and more diverse datasets, particularly including rarer BCa types such as ILC and DCIS, are needed to further improve and evaluate the models. In addition, future work should focus on improving model efficiency for fast intraoperative decision-making, as well as on further enhancing model trustworthiness to support effective collaboration with physicians. Moreover, the clinical usability and value of the proposed DL models should be investigated. Overall, the work presented in this dissertation is a first step towards a DL-based decision-support system to enhance IMA in BCa using micro-PET-CT, facilitating the adoption of this technique in clinical practice.

Samenvatting

Patiënten met borstkanker in een vroeg stadium worden standaard behandeld met borstsparende chirurgie, gevolgd door radiotherapie. Een succesvolle chirurgische ingreep vereist volledige verwijdering van de tumor met een marge van omliggend gezond weefsel, resulterend in negatieve snijranden. Positieve snijranden, die wijzen op resterende tumorcellen aan de rand van het resectiespecimen, vormen de belangrijkste risicofactor voor lokaal recidief na borstsparende chirurgie. De gouden standaard voor margebeoordeling is histopathologisch onderzoek van coupes van het resectiespecimen. Volgens de SSO-ASTRO ("Society of Surgical Oncology and American Society for Radiation Oncology") consensus van 2014 wordt voor invasief carcinoom van het niet-speciale type (NST) en invasief lobulair carcinoom (ILC) een tumorvrije marge van meer dan 0 mm als voldoende beschouwd, terwijl voor ductaal carcinoom in situ (DCIS) een minimale marge van 2 mm wordt aanbevolen. Desondanks worden bij ongeveer 20% van de patiënten na de eerste operatie positieve snijranden vastgesteld, hoewel in de literatuur grote variaties in dit percentage worden gerapporteerd. Deze patiënten moeten vaak een heroperatie ondergaan om resterende tumorcellen te verwijderen uit de borst, wat gepaard gaat met psychologische, fysieke en economische belasting. Snelle en betrouwbare intraoperatieve margebeoordeling is daarom essentieel om heroperaties te voorkomen. Bestaande technieken, zoals vriescoupes, imprint cytologie, intraoperatieve echografie en specimenradiografie, kennen echter belangrijke beperkingen. Momenteel ontbreekt een kosteneffectieve methode die voldoende nauwkeurig is en de chirurgische workflow niet verstoort.

Recent is een veelbelovende techniek ontwikkeld voor intraoperatieve margebeoordeling: *ex vivo* beeldvorming van resectiespecimens met hoge-resolutie (micro-) positronemissietomografie (PET), gecombineerd met computertomografie (CT). Een compacte, mobiele micro-PET-CT scanner kan meteen in de operatiekamer 3D beelden genereren van verwijderde borstkankerspecimens, waardoor zowel de status als de breedte van de snijranden gedetailleerd kan worden beoordeeld. Klinische studies hebben de haalbaarheid en veiligheid van deze techniek aangetoond en lieten zien dat micro-PET-CT een significant hoger succespercentage behaalt bij het opsporen en adequaat aanpakken van positieve snijranden in borstkankerspecimens dan standaard intraoperatieve technieken. In deze studies werden beelden handmatig geïnterpreteerd door artsen in ongeveer één minuut per 3D beeld. Hoewel gezamenlijke beoordeling door meerdere artsen de reproduceerbaarheid verbeterde, bleef interwaarnemer-variantie bestaan. Dit proefschrift onderzoekt daarom de eerste stappen in de ontwikkeling van een beslissingsondersteunend systeem dat artsen kan helpen bij de interpretatie van micro-PET-CT beel-

den, met als doel de interpretatiesnelheid, de interwaarnemer-variatie en mogelijks de diagnostische prestaties verder te verbeteren. Daartoe is een deep learning (DL) model ontwikkeld voor automatische segmentatie van borsttumoren in micro-PET-CT beelden van borstkankerspecimens, waarmee tevens de margestatus kan worden voorspeld.

Voor de ontwikkeling van dit model werd eerst een dataset samengesteld van micro-PET-CT beelden van borstkankerspecimens met bijbehorende tumorsegmentaties op basis van histopathologie. Hiervoor ontwikkelden we een nieuwe methode om micro-PET-CT beelden van chirurgische resectiespecimens automatisch en nauwkeurig te coregistreren met overeenkomstige digitale beelden van histopathologische coupes. Deze methode combineert een beeldacquisitieprotocol dat weefselvorming minimaliseert met een wiskundig coregistratie-algoritme dat de alignering van de beelden verder optimaliseert. Het acquisitieprotocol omvat drie stappen: (1) acquisitie van micro-PET-CT beelden van lamellen van verse resectiespecimens met een dikte van ongeveer 2 mm, ingesloten in weefselcassettes; (2) acquisitie van micro-CT beelden van dezelfde lamellen na fixatie in formaline om weefselvorming in kaart te brengen; en (3) acquisitie van een digitaal beeld van één coupe met hematoxyline-eosine (H&E) kleuring per lamel. Vervolgens werd een automatisch coregistratie-algoritme toegepast om de micro-PET-CT beelden van de verse lamellen te aligneren met de digitale coupes, gestuurd door de micro-CT beelden van de gefixeerde lamellen. In vergelijking met bestaande methoden voor het coregistreren van *in vivo* PET beelden met histopathologie is deze aanpak minder arbeidsintensief en tijdrovend. Naast de standaard histopathologische verwerking is enkel het snijden van verse lamellen vereist, waarna een volledig automatische coregistratie wordt uitgevoerd aan de hand van intensiteitsgebaseerde gelijkenismetrieken.

De coregistratie-methode werd toegepast op respectievelijk 53, 19 en 10 cassettes van 19 NST, 6 ILC en 3 DCIS specimens van patiënten die preoperatief 0.8 MBq/kg [^{18}F]fluorodeoxyglucose ([^{18}F]FDG) kregen toegediend. Hoewel de cassettes weefselvorming tot een minimum beperkten, werd de beste alignering van de micro-PET-CT beelden met de digitale coupes bereikt na deformeerbare coregistratie. Dit wijst op de noodzaak om te corrigeren voor niet-rigide weefselvormingen. Met microcalcificaties als referentiepunten bedroeg de uiteindelijke gemiddelde ruimtelijke coregistratiefout tussen micro-PET en histopathologie 0.69 ± 0.23 mm, wat vergelijkbaar is met de ruimtelijke resolutie van de micro-PET beelden. De verkregen 2D gecoregistreerde beelden maakten het mogelijk om de SUVs ("standardised uptake values") van verschillende borstweefsels te analyseren. De borstweefsels werden geïdentificeerd op de digitale coupes door een ervaren borstpatholoog. Doorgaans vertoonde (pre)maligne weefsel (waaronder NST, ILC en DCIS) hogere SUVs dan gezond weefsel (waaronder normaal klier-, bind- en vetweefsel). Dit suggereert dat

ILC en DCIS tumoren met micro-PET-CT kunnen worden gevisualiseerd, terwijl ze doorgaans moeilijker detecteerbaar zijn met *in vivo* PET. Zoals verwacht vertoonden ook inflammatie en huid een verhoogde opname van [^{18}F]FDG. Verder konden tumor annotaties op basis van histopathologie worden overgedragen naar de gealigneerde 2D micro-PET-CT beelden om referentiesegmentaties van borsttumoren te verkrijgen, essentieel voor de ontwikkeling en validatie van segmentatiemodellen gebaseerd op DL. Hoewel deze methode momenteel enkel werd toegepast op (sterk vervormbare) borstkankerspecimens, is zij naar verwachting breder toepasbaar op diverse resectiespecimens en radiotracers. Daarmee kan inzicht worden verkregen in de distributie van (nieuwe) radiotracers in humane maligniteiten met een resolutie op submillimeter niveau.

Vervolgens gebruikten we de verzamelde dataset van 2D micro-PET-CT beelden, voorzien van op histopathologie gebaseerde referentiesegmentaties, om een DL model te ontwikkelen voor automatische segmentatie van borsttumoren. Hiervoor maakten we gebruik van een gestratificeerd, genest vijfvoudig groeps kruisvalidatieschema om een ensemble van 2D ResU-Nets (“residual U-Nets”) te trainen en te evalueren voor het segmenteren van (pre)maligne weefsel in micro-PET-CT beelden van borstkankerspecimens. De modellen werden eerst getraind op de grootste groep van NST cases, waarna transfer learning werd toegepast op de gecombineerde groep van ILC en DCIS cases. Voor de segmentatie van NST tumoren behaalden de ResU-Nets een DSC (“Dice similarity coefficient”) van 0.73 ± 0.20 , en een PR-AUC (“area under the curve of the precision-recall curve”) van 0.83 ± 0.14 . We vergeleken de prestaties van de ResU-Nets met conventionele PET intensiteitsdrempels en stelden vast dat de ResU-Nets voor de NST cases significant beter presteerden wat betreft de PR-AUC. De segmentatie van ILC en DCIS tumoren bleek uitdagender, met DSC-waarden van respectievelijk 0.49 ± 0.26 en 0.03 ± 0.02 , en PR-AUC-waarden van 0.60 ± 0.25 en 0.03 ± 0.03 . Hoewel de ResU-Nets betere prestaties lieten zien dan de PET intensiteitsdrempels voor deze subtypes, konden er geen statistisch significante verbeteringen worden aangetoond door de beperkte grootte van de dataset. Daarnaast toonden we aan dat de combinatie van metabole informatie uit micro-PET met hoge-resolutie structurele informatie uit micro-CT essentieel is voor nauwkeurige tumorsegmentatie, waarbij modellen met een multimodale input beter presteerden dan modellen met één enkele modaliteit als input. Een uniek aspect van deze studie is het gebruik van histopathologie als referentiestandaard voor tumorsegmentatie, in plaats van de gebruikelijke handmatige annotaties door experts, wat resulteert in een klinisch relevantere representatie van de tumoromvang.

In een volgende stap werd geëvalueerd in welke mate de ResU-Nets in staat zijn de margestatus van borstkankerspecimens accuraat te voorspellen. Hiervoor werden de datasets gebruikt die in de BIMAP en BrIMA studies werden verzameld, met 3D micro-PET-CT beelden en de referentiemarge-

status volgens histopathologie. De BIMAP studie verzamelde beelden van 41 borstkankerspecimens in het Universitair Ziekenhuis Gent, terwijl de BrIMA studie 152 specimens in beeld bracht uit zes ziekenhuizen in drie Europese landen. Beide studies includeerden patiënten met borstkanker in een vroeg stadium (waaronder NST, ILC, DCIS en neoadjuvant behandelde (NAT) borstkanker), waarbij NST het meest voorkomende subtype was met respectievelijk 31 en 84 patiënten in de BIMAP en BrIMA studies. Het ensemble van ResU-Nets werd toegepast op deze voorheen ongeziene beelden om (pre)maligne weefsel te segmenteren, waarna de margestatus als positief werd voorspeld indien de gesegmenteerde tumor de rand van het specimen raakte. De voorspelde margestatus werd vergeleken met de werkelijke histopathologische margestatus. Voor de groep van NST patiënten behaalden de ResU-Nets F1-scores van respectievelijk 84% en 41% voor de BIMAP en BrIMA datasets. Voor de gecombineerde groep van alle borstkanker subtypes bedroegen de F1-scores respectievelijk 76% en 50%. De margestatus voorspellingen van de ResU-Nets werden ook vergeleken met de voorspellingen van conventionele PET intensiteitsdrempels, evenals met intraoperatieve margebeoordelingen door artsen (BrIMA), en postoperatieve meerderheidsbesluiten van respectievelijk zeven en vier artsen (BIMAP en BrIMA). Zowel voor de NST tumoren als voor de gecombineerde groep van alle borstkanker subtypes in de BIMAP en BrIMA datasets presteerden de ResU-Nets statistisch gelijkwaardig aan de intraoperatieve en postoperatieve beslissingen van de artsen, terwijl ze consistent beter presteerden dan PET intensiteitsdrempels. Belangrijk is dat de ResU-Nets prestaties op artsenniveau behaalden voor een grote multicentrische dataset met diverse borstkanker subtypes, representatief voor de klinische praktijk. Hoewel voor kleinere subgroepen (ILC, DCIS en NAT) niet altijd statistische significantie kon worden aangetoond, vertoonden de ResU-Nets doorgaans hogere specificiteit, precisie en F1-scores dan de PET intensiteitsdrempels. Bovendien presteerden de modellen, ondanks dat ze niet waren getraind op NAT tumoren, ook voor deze subgroep vergelijkbaar met artsen, wat wijst op hun robuustheid.

Tot slot vereist veilige klinische implementatie van dergelijke DL modellen inzicht in de onzekerheid van de voorspellingen, iets wat niet wordt weergegeven in binaire tumorsegmentaties of margestatus labels. De tumorprobabiliteitsmappen voorspeld door de ResU-Nets drukken deze onzekerheid uit op pixelniveau. Om de betrouwbaarheid en interpreteerbaarheid van de voorspellingen te verbeteren, werden de tumorprobabiliteitsmappen gekalibreerd met behulp van de inductieve cross-Venn-ABERS methode, met focus op de grootste subgroep van patiënten met NST borstkanker. De kalibratie werd uitgevoerd en geëvalueerd op de dataset van 53 2D micro-PET-CT beelden van lamellen van 19 NST borstkankerspecimens, met de histopathologische tumorsegmentaties als referentiestandaard. De validatie sets uit het genest kruisvalidatieschema, gebruikt om de segmentatiemodellen te trainen, werden gebruikt als kalibratie sets. Kalibratie verbeterde de overeenkomst tussen voorspelde

tumorkansen op pixelniveau en geobserveerde frequenties van kwaadaardige pixels aanzienlijk. De ECE (“expected calibration error”), BCE (“binary cross-entropy”) loss en Brier loss werden met respectievelijk 60.5%, 17.7% en 14.8% verminderd tot eindwaarden van 0.017, 0.172 en 0.052. De prestaties voor tumorsegmentatie, geëvalueerd op basis van ROC-AUC (“area under the curve of the receiver operating characteristic”) en PR-AUC, bleven na kalibratie stabiel. Er werden ook gekalibreerde tumorprobabiliteitsmappen gegenereerd voor de 84 3D micro-PET-CT beelden van NST borstkankerspecimens in de BrIMA dataset, voorzien van de referentiemargestatus volgens histopathologie. Kalibratie leidde tot een lichte afname in PR-AUC voor margebeoordeling. Kwalitatief boden de tumorprobabiliteitsmappen transparante en ruimtelijk gedetailleerde visualisaties van de voorspellingsonzekerheid, en maakte kalibratie deze schattingen betrouwbaarder, wat een betere identificatie van onzekere tumorgrenzen en ambigue margegebieden mogelijk maakte. Deze bevindingen suggereren dat gekalibreerde tumorprobabiliteitsmappen kunnen bijdragen aan beter geïnformeerde, efficiëntere en risicobewuste klinische besluitvorming.

Samengevat zijn de voorgestelde DL modellen in staat om zowel de tumorsegmentatie als de margestatus te voorspellen op basis van een micro-PET-CT beeld, en om ruimtelijk gedetailleerde voorspellingsonzekerheid weer te geven. Deze modellen hebben het potentieel om de consistentie te verhogen en de besluitvorming te versnellen bij de interpretatie van micro-PET-CT beelden van borstkankerspecimens. Hoewel de resultaten veelbelovend zijn, is verdere validatie noodzakelijk voordat de DL modellen in de klinische praktijk kunnen worden toegepast. Om de modellen verder te verbeteren en evalueren zijn grotere en meer diverse datasets nodig, in het bijzonder met zeldzamere borstkanker subtypes zoals ILC en DCIS. Daarnaast dient toekomstig onderzoek zich te richten op de ontwikkeling van efficiëntere modellen voor snelle intraoperatieve besluitvorming, evenals op het verder verbeteren van de betrouwbaarheid van de modellen om de samenwerking met artsen te optimaliseren. Verder moeten de klinische bruikbaarheid en waarde van de voorgestelde DL modellen worden onderzocht. Kortom, dit proefschrift vormt een eerste stap richting een DL-gebaseerd beslissingsondersteunend systeem voor intraoperatieve margebeoordeling op basis van micro-PET-CT beelden van borstkankerspecimens, waardoor de toepassing van deze techniek in de klinische praktijk wordt vergemakkelijkt.

List of Acronyms

[¹⁸ F]FDG	[¹⁸ F]fluorodeoxyglucose
1mmCD	contour Dice with a 1 mm tolerance
95HD	95 th percentile of the Hausdorff distance
AI	artificial intelligence
ANN	artificial neural network
ASGD	adaptive stochastic gradient descent
AUC	area under the curve
BCa	breast cancer
BCE	binary cross-entropy
BCS	breast-conserving surgery
CE	Conformité Européenne (European Conformity)
CNN	convolutional neural network
CT	computed tomography
DCIS	ductal carcinoma in situ
DL	deep learning
DSC	Dice similarity coefficient
ECE	expected calibration error
ER	oestrogen receptor
FBP	filtered backprojection
FN	false negative
FP	false positive
FWHM	full width at half maximum
GAN	generative adversarial network
GD	gradient descent
GDPR	General Data Protection Regulation
H&E	haematoxylin and eosin
HER2	human epidermal growth factor receptor 2
HR	hormone receptor
HU	Hounsfield unit
ICC	intraclass correlation coefficient
ILC	invasive lobular carcinoma
ILC-AI	ILC pretreated with aromatase inhibitors
IMA	intraoperative margin assessment

ISRA	image space reconstruction algorithm
IVDR	In Vitro Diagnostic Regulation
LBM	lean body mass
LCIS	lobular carcinoma in situ
LOO	leave-one-out
LOR	line of response
ML	machine learning
MLEM	maximum likelihood expectation maximisation
MLP	multilayer perceptron
MRI	magnetic resonance imaging
NAC	neoadjuvant chemotherapy
NAT	neoadjuvant-treated
NCC	normalised correlation coefficient
NMI	normalised mutual information
NST	no special type
NST-AI	NST pretreated with aromatase inhibitors
OSEM	ordered subset expectation maximisation
PET	positron emission tomography
PMT	photomultiplier tube
PR	precision-recall
PR-AUC	area under the curve of the precision-recall curve
PreLU	parametric ReLU
ReLU	rectified linear unit
ResNet	Residual Network
ResU-Net	Residual U-Net
RMS	root mean square
ROC	receiver operating characteristic
ROC-AUC	area under the curve of the receiver operating characteristic
RSS	root sum square
SaMD	Software as a Medical Device
SERM	selective oestrogen receptor modulator
SGD	stochastic gradient descent
SiPM	silicon photomultiplier
SOC	standard-of-care
SSO-ASTRO	Society of Surgical Oncology and American Society for Radiation Oncology
SUV	standardised uptake value
TDLU	terminal duct lobular unit

TN	true negative
TP	true positive
US	ultrasound
WSI	whole slide image

List of Publications

Journal Papers

Jens M. Debacker, **Luna Maris**, Fleur Cordier, David Creytens, Philippe Deron, Benedicte Descamps, Yves D'Asseler, Kathia De Man, Vincent Keereman, Sasha Libbrecht, Vanessa Schelfhout, Koen Van De Vijver, Christian Vanhove, and Wouter Huvenne. Direct co-registration of [18F]FDG uptake and histopathology in surgically excised malignancies of the head and neck: a feasibility study. *European Journal of Nuclear Medicine and Molecular Imaging*, 50(7):2127–2139, 2023.

Luna Maris, Menekse Göker, Jens M. Debacker, Kathia De Man, Bliede Van den Broeck, Jo Van Dorpe, Koen Van de Vijver, Vincent Keereman, and Christian Vanhove. Method for co-registration of high-resolution specimen PET-CT with histopathology to improve insight into radiotracer distributions. *European Journal of Nuclear Medicine and Molecular Imaging Physics*, 11(1):85, 2024.

Luna Maris, Menekse Göker, Kathia De Man, Bliede Van den Broeck, Sofie Van Hoecke, Koen Van de Vijver, Christian Vanhove, and Vincent Keereman. Supporting intraoperative margin assessment using deep learning for automatic tumour segmentation in breast lumpectomy micro-PET-CT. *npj Breast Cancer*, Nature Publishing Group, 11(1):88, 2025.

Patent Applications

Luna Maris, Vincent Keereman, and Christian Vanhove. Method for co-registration of nuclear images with microscopic histopathology images, Patent with International Application Number PCT/EP2025/054989, Publication Number WO/2025/186049, and Priority Date 08.03.2024.

Conference Contributions

Luna Maris, Jens M. Debacker, Kathia De Man, Menekse Göker, Vincent Keereman, Bliede Van den Broeck, Koen Van de Vijver, Roel Van Holen, and Christian Vanhove. Method for direct correlation of high-resolution ¹⁸F-FDG-PET-CT images with histopathology in resected breast cancer specimens. Abstract and poster presentation. *European Association of Nuclear Medicine*, Barcelona, Spain, 2022.

Luna Maris, Kathia De Man, Filip Gryspeerdt, Anne Hoorens, Vincent Keereman, Anneleen Kiekens, Bliede Van den Broeck, Leen Van de Sande, Christian Vanhove, and Frederik Berrevoet. ^{18}F -FDG-PET-CT specimen imaging for perioperative visualization of pancreatic adenocarcinoma: a proof-of-concept study. Abstract and poster presentation. *European Society of Surgical Oncology*, Bordeaux, France, 2022.

Luna Maris, Menekse Göker, Kathia De Man, Bliede Van den Broeck, Roel Van Holen, Vincent Keereman, Koen Van de Vijver, and Christian Vanhove. Standardized uptake values of breast cancer types in sub-millimetre ^{18}F -FDG-PET-CT: direct correlation with histopathology. Abstract and poster presentation. *European Association of Nuclear Medicine*, Vienna, Austria, 2023.

Luna Maris, Menekse Göker, Kathia De Man, Bliede Van den Broeck, Vincent Keereman, Koen Van de Vijver, and Christian Vanhove. Can micro-PET-CT imaging of surgical specimens reduce sampling error? Research in breast cancer. Abstract and poster presentation. *Belgian Week of Pathology*, Brussels, Belgium, 2023.

Luna Maris, Menekse Göker, Kathia De Man, Bliede Van den Broeck, Jo Van Dorpe, Koen Van de Vijver, Vincent Keereman, and Christian Vanhove. Deep learning model for tumor segmentation in micro-PET-CT images of breast cancer specimens. Abstract and poster presentation. *IEEE Medical Imaging Congress*, Tampa, Florida, 2024.

Luna Maris, Menekse Göker, Kathia De Man, Bliede Van den Broeck, Koen Van de Vijver, Christian Vanhove, and Vincent Keereman. Deep learning model to automatically assess margins of lumpectomy specimens using micro-PET-CT images. Abstract and poster presentation. *Ghent Pathology*, Ghent, Belgium, 2025.

Luna Maris, Menekse Göker, Kathia De Man, Bliede Van den Broeck, Sofie Van Hoecke, Koen Van de Vijver, Christian Vanhove, and Vincent Keereman. Deep learning model for automatic tumour segmentation and margin status prediction in micro-PET-CT images of breast lumpectomy specimens. Abstract and oral presentation. *European Association of Nuclear Medicine*, Barcelona, Spain, 2025.

Luna Maris, Menekse Göker, Kathia De Man, Bliede Van den Broeck, Sofie Van Hoecke, Koen Van de Vijver, Christian Vanhove, and Vincent Keereman. Deep learning-based margin assessment in micro-PET-CT images of breast lumpectomy specimens: validation on a multicentre dataset. Abstract, poster, and oral presentation. *AI in Medicine*, Krakow, Poland, 2026.

Table of Contents

Acknowledgements	i
Summary	v
Samenvatting	xi
List of Acronyms	xvii
List of Publications	xxi
Table of Contents	xxiii
List of Figures	xxix
List of Tables	xxxv
1. Introduction	1
1.1. Clinical rationale	1
1.2. Research objectives	2
1.3. Outline	3
 I. Clinical and Technical Background	 5
2. Micro-PET-CT specimen imaging for intraoperative margin assessment in breast-conserving surgery	7
2.1. Introduction	7
2.2. Breast cancer	8
2.2.1. Breast anatomy	8
2.2.2. Breast pathology	9
2.2.2.1. Benign conditions	9
2.2.2.2. Premalignant conditions	10
2.2.2.3. Malignant conditions	11
2.2.3. Breast cancer classification	11
2.2.3.1. Histological classification	11
2.2.3.2. Molecular classification	13
2.2.3.3. Grading	13
2.2.3.4. Staging	14
2.2.4. Epidemiology	14
2.2.4.1. Incidence	14
2.2.4.2. Risk factors	15
2.2.4.3. Prevalence of breast cancer subtypes	15
2.2.5. Symptoms	16

2.2.6.	Diagnosis	16
2.2.6.1.	Physical examination	16
2.2.6.2.	Medical imaging	17
2.2.6.3.	Biopsy and histopathological assessment	19
2.2.7.	Treatment	19
2.2.7.1.	Local treatment: Surgery	19
2.2.7.2.	Locoregional treatment: Radiotherapy	20
2.2.7.3.	Systemic treatment	20
2.3.	Breast-conserving surgery (BCS) and positive margins	22
2.3.1.	Principles of successful BCS	22
2.3.2.	Margin status in BCS	22
2.3.3.	Histopathology as gold standard for margin assessment	23
2.3.4.	Limitations of histopathology	24
2.3.5.	Prevalence and impact of positive margins	26
2.4.	Intraoperative margin assessment (IMA)	26
2.4.1.	Standard-of-care (SOC) techniques	27
2.4.2.	Emerging imaging-based techniques	28
2.5.	Micro-PET-CT specimen imaging as IMA technique	30
2.5.1.	Fundamental principles of PET-CT	31
2.5.1.1.	Computed tomography (CT)	31
2.5.1.2.	Positron emission tomography (PET)	35
2.5.2.	Rationale and system design for ex vivo micro-PET-CT	39
2.5.3.	Feasibility and safety of intraoperative implementation	41
2.5.4.	Diagnostic performance for IMA in breast cancer	42
2.5.4.1.	BIMAP	42
2.5.4.2.	BrIMA	44
3. Fundamentals of deep learning and its application to medical image analysis		47
3.1.	Introduction	47
3.2.	Fundamentals of deep learning	48
3.2.1.	Artificial intelligence, machine learning, and deep learning	48
3.2.2.	Supervised, unsupervised, and reinforcement learning	48
3.2.3.	Model architecture: Deep artificial neural networks	49
3.2.4.	Model training	51
3.2.5.	Model generalisation and regularisation	52
3.2.5.1.	Data splits and cross-validation	52
3.2.5.2.	Bias-variance trade-off	53
3.2.5.3.	Regularisation techniques	54
3.3.	Deep learning for computer vision in medical images	56
3.3.1.	Classification and segmentation	56
3.3.2.	Convolutional neural networks	57
3.3.2.1.	Different layer types	58
3.3.2.2.	ResNet for classification	60
3.3.2.3.	U-Net for segmentation	60

3.3.3.	Performance evaluation metrics	64
3.3.3.1.	Classification	64
3.3.3.2.	Segmentation	66
3.3.4.	Challenges related to deep learning for medical image analysis	67
3.3.4.1.	Scarcity of well-curated annotated datasets	67
3.3.4.2.	Class imbalance	68
3.3.4.3.	High-dimensional complex data	68
3.3.4.4.	Data heterogeneity and domain shifts	68
3.3.5.	Emerging trends in deep learning for medical image analysis	69
3.3.5.1.	Data-Efficient and collaborative learning strategies	69
3.3.5.2.	Architectural innovations	70
3.3.5.3.	Advanced optimisation and supervision techniques	71
3.3.5.4.	Techniques for efficient and trustworthy deployment	72
3.4.	Ground-truth generation through medical image co-registration	73

II. Research Results

77

4. Co-registration of micro-PET-CT with histopathology images of lumpectomy specimens to obtain ground-truth tumour segmentations

79

4.1.	Introduction	81
4.2.	Methods	83
4.2.1.	Co-registration method	83
4.2.1.1.	Image acquisition scheme	83
4.2.1.2.	Mathematical co-registration algorithm	85
4.2.2.	Application and evaluation in breast cancer	88
4.2.2.1.	Study design and subjects	88
4.2.2.2.	Image acquisition	88
4.2.2.3.	Mathematical co-registration	90
4.2.2.4.	Co-registration evaluation	90
4.2.2.5.	Standardized uptake values	91
4.3.	Results	93
4.3.1.	Image acquisition in breast cancer	93
4.3.2.	Mathematical co-registration in breast cancer	93
4.3.3.	Standardised uptake values in breast cancer	98
4.4.	Discussion	101
4.5.	Conclusions	104

5. Breast tumour segmentation in micro-PET-CT images of lumpectomy specimens	105
5.1. Introduction	107
5.2. Methods	108
5.2.1. Dataset	108
5.2.1.1. Study design and subjects	108
5.2.1.2. Micro-PET-CT images of breast lamellas	108
5.2.1.3. Breast tumour segmentation labels	109
5.2.2. Deep learning model	110
5.2.2.1. Architecture: ResU-Net	110
5.2.2.2. Data splits, preprocessing, and augmentation	110
5.2.2.3. Hyperparameter optimisation	111
5.2.2.4. Training	112
5.2.3. Intensity thresholding	112
5.2.4. Segmentation performance evaluation	112
5.3. Results	114
5.3.1. Dataset	114
5.3.2. Trained models	114
5.3.3. Segmentation performance	115
5.4. Discussion	120
5.5. Conclusions	122
6. Margin status prediction based on micro-PET-CT images of lumpectomy specimens	123
6.1. Introduction	125
6.2. Methods	125
6.2.1. Datasets	125
6.2.1.1. Study design and subjects: BIMAP and BrIMA	126
6.2.1.2. Micro-PET-CT images of breast cancer specimens	126
6.2.1.3. Margin status classification labels	127
6.2.2. Workflow for margin status prediction	127
6.2.2.1. Data preprocessing and splits	128
6.2.2.2. Deep learning model	128
6.2.2.3. Intensity thresholding	131
6.2.3. Margin assessment performance evaluation	131
6.3. Results	134
6.3.1. Margin assessment performance	134
6.3.1.1. BIMAP	134
6.3.1.2. BrIMA	138
6.3.2. Performance stability across operating points	142
6.4. Discussion	146
6.5. Conclusions	149
7. Calibration of tumour probability maps for trustworthy interpretation of micro-PET-CT images of lumpectomy specimens	151
7.1. Introduction	153

7.2.	Methods	154
7.2.1.	Datasets	154
7.2.1.1.	2D micro-PET-CT of lamellas with tumour segmentation labels	154
7.2.1.2.	3D micro-PET-CT of specimens with margin classification labels	154
7.2.2.	Venn-ABERS calibration of pixel-level tumour probabilities	155
7.2.3.	Calibration evaluation	156
7.2.3.1.	Quantitative calibration assessment	156
7.2.3.2.	Impact on segmentation and margin assessment performance	159
7.2.3.3.	Qualitative analysis of calibrated tumour probability maps	159
7.3.	Results	160
7.3.1.	Quantitative calibration assessment	160
7.3.2.	Impact on segmentation and margin assessment performance	162
7.3.3.	Qualitative analysis of calibrated tumour probability maps	163
7.4.	Discussion	169
7.5.	Conclusions	171

III. General Conclusions and Future Perspectives 173

8. Conclusions and future perspectives 175

8.1.	Results summary	175
8.2.	Future perspectives	177
8.2.1.	Data and ground truth development	177
8.2.1.1.	Improved co-registration using high-contrast images	177
8.2.1.2.	3D micro-PET-CT data with ground-truth tumour segmentations	177
8.2.1.3.	Improved insight into radiotracer uptake in breast tumours	178
8.2.2.	Improving model performance	178
8.2.2.1.	Data-efficient and collaborative learning strategies	178
8.2.2.2.	Alternative model architectures	179
8.2.2.3.	Improved optimisation and supervision	180
8.2.3.	Trustworthy and efficient deployment in clinical practice	180
8.2.4.	Regulatory considerations	181
8.2.5.	Extension towards other cancer types and radiotracers	181
8.3.	Conclusions	182

- IV. Appendices** **185**
- A. Co-registration parameters** **187**
- B. Impact of fixated micro-CT guidance on co-registration accuracy** **189**
- C. ResU-Net hyperparameters and learning curves** **191**
- D. Tumour segmentation performance for micro-PET-CT images of ILC and DCIS lumpectomy specimens** **195**
- E. Margin assessment performance for micro-PET-CT images of ILC, DCIS, and neoadjuvant-treated lumpectomy specimens** **203**
 - E.1. BIMAP 203
 - E.2. BrIMA 210
- F. Leave-one-out strategy to choose an operating point for margin status prediction from micro-PET-CT images of lumpectomy specimens** **217**
 - F.1. BIMAP 218
 - F.2. BrIMA 221
- References** **225**

List of Figures

2.1.	Anatomy of the female breast.	9
2.2.	Histological breast cancer subtypes.	12
2.3.	Illustration of breast-conserving surgery or lumpectomy and the associated surgical margin status.	23
2.4.	Histopathological processing of a breast cancer specimen.	25
2.5.	Electromagnetic spectrum.	30
2.6.	Basic principles of CT imaging.	34
2.7.	Basic principles of PET imaging.	39
2.8.	Micro-PET image of a Derenzo phantom filled with [¹⁸ F]fluorodeoxyglucose ([¹⁸ F]FDG) acquired with the AURA [®] 10 from XEOS.	40
2.9.	Two examples of micro-PET-CT images of breast cancer specimens.	41
2.10.	Workflow for intraoperative micro-PET-CT imaging of lumpectomy specimens.	42
3.1.	Relation between artificial intelligence, machine learning, and deep learning.	48
3.2.	The perceptron as the basic building block of an artificial neural network.	50
3.3.	Fully connected artificial neural network or multilayer perceptron.	50
3.4.	Iterative training of an artificial neural network (ANN).	52
3.5.	Nested five-fold cross-validation.	53
3.6.	Bias-variance trade-off when choosing a decision boundary for a binary classification problem.	54
3.7.	Early stopping.	55
3.8.	2D convolution operation.	57
3.9.	Illustration of various normalisation methods used in convolutional neural networks: batch, layer, instance, and group normalisation.	59
3.10.	2D Residual Network (ResNet) for image classification	62
3.11.	2D U-Net architecture for image segmentation.	63
3.12.	Contingency table.	65
3.13.	Receiver operating characteristic (ROC) and precision-recall (PR) curve.	66
3.14.	Iterative image co-registration.	75
4.1.	Illustration of the proposed image acquisition scheme for a breast cancer specimen.	84
4.2.	Illustration of the four-step co-registration algorithm.	87
4.3.	Overview of images collected for one breast lamella.	89

- 4.4. Illustration of the Gaussian filtering process of a tissue annotation. 92
- 4.5. Intermediate results of the mathematical co-registration process for one breast lamella. 95
- 4.6. Overview of the alignment between the micro-PET-CT images and histopathology whole slide images (WSIs) after mathematical co-registration. 96
- 4.7. Illustration of corresponding microcalcifications in micro-CT images and histopathology whole slide images (WSIs). 97
- 4.8. Box plots showing the spread in standardised uptake values (SUVs) of different breast tissue types. 99
- 4.9. Box plots showing the spread across patients of ratios of standardised uptake values (SUVs) for different breast tissue types relative to healthy tissue. 100

- 5.1. Illustration of the Residual U-Net (ResU-Net) architecture used in this work. 111
- 5.2. Example of a micro-PET-CT slice with the corresponding true tumour segmentation, as annotated on the histopathological whole slide image (WSI) by the pathologist, for an invasive carcinoma of no special type (NST). 114
- 5.3. Example of the tumour segmentations predicted with intensity thresholding and Residual U-Net (ResU-Net) for an invasive carcinoma of no special type. 118
- 5.4. Box plots showing metric distributions assessing agreement between true and predicted tumour segmentations for test patients with invasive carcinoma of no special type. 119

- 6.1. Illustration of the automated workflow for tumour segmentation and margin status prediction from a micro-PET-CT image of a lumpectomy specimen using a deep learning (DL) model. 130
- 6.2. Predicted tumour contours for three invasive carcinomas of no special type from the BIMAP dataset. 136
- 6.3. Contingency tables for margin status prediction in the BIMAP dataset, shown separately for invasive carcinomas of no special type (NST), and for all breast cancer (BCa) subtypes combined. 137
- 6.4. P-values (p) resulting from McNemar’s test to compare different approaches to predict the margin status for the BIMAP dataset, shown separately for invasive carcinomas of no special type (NST), and for all breast cancer (BCa) subtypes combined. 137
- 6.5. Predicted tumour contours for three invasive carcinomas of no special type from the BrIMA dataset. 140

6.6.	Contingency tables for margin status prediction in the BrIMA dataset, shown separately for invasive carcinomas of no special type (NST), and for all breast cancer (BCa) subtypes combined.	141
6.7.	P-values (p) resulting from McNemar's test to compare different approaches to predict the margin status for the BrIMA dataset, shown separately for invasive carcinomas of no special type (NST), and for all breast cancer (BCa) subtypes combined.	141
6.8.	Receiver operating characteristic (ROC) and precision-recall (PR) curves for margin status prediction for the BIMAP dataset, shown separately for invasive carcinomas of no special type (NST), and for the combined group of all breast cancer (BCa) subtypes.	144
6.9.	Receiver operating characteristic (ROC) and precision-recall (PR) curves for margin status prediction for the BrIMA dataset, shown separately for invasive carcinomas of no special type (NST), and for the combined group of all breast cancer (BCa) subtypes.	145
7.1.	Illustration of inductive Venn-ABERS calibration for a single test pixel using one calibration fold.	158
7.2.	Distribution and calibration plots for pixel-level tumour probabilities before and after Venn-ABERS calibration.	161
7.3.	Visual comparison of tumour probability maps for lumpectomy lamellas before and after Venn-ABERS calibration.	164
7.4.	Visual comparison of tumour probability maps before and after Venn-ABERS calibration for a BrIMA lumpectomy specimen of invasive carcinoma of no special type with a positive margin according to histopathology.	165
7.5.	Visual comparison of tumour probability maps before and after Venn-ABERS calibration for a BrIMA lumpectomy specimen of invasive carcinoma of no special type with a negative margin according to histopathology.	166
7.6.	Visual comparison of tumour probability maps before and after Venn-ABERS calibration for a BrIMA lumpectomy specimen of invasive carcinoma of no special type with a negative margin according to histopathology.	167
7.7.	Visual comparison of tumour probability maps before and after Venn-ABERS calibration for a BrIMA lumpectomy specimen of invasive carcinoma of no special type with a positive margin according to histopathology.	168
B.1.	Difference between micro-CT images of fresh lamellas co-registered with the histopathology whole slide images with and without guidance of the micro-CT of the fixated lamellas.	190

C.1. Evolution of training and validation Dice loss across epochs during the training of 75 Residual U-Nets (ResU-Nets) on data from cases with invasive carcinoma of no special type (NST) 191

C.2. Evolution of training and validation Dice loss across epochs during the transfer learning of 75 Residual U-Nets (ResU-Nets) on data from cases with invasive lobular carcinoma (ILC) and ductal carcinoma in situ (DCIS). 192

D.1. Example of the tumour segmentations predicted with intensity thresholding and Residual U-Net (ResU-Net) for an invasive lobular carcinoma. 196

D.2. Example of the tumour segmentations predicted with intensity thresholding and Residual U-Net (ResU-Net) for a ductal carcinoma in situ. 197

D.3. Box plots showing metric distributions assessing agreement between true and predicted tumour segmentations for test patients with invasive lobular carcinoma. 198

D.4. Box plots showing metric distributions assessing agreement between true and predicted tumour segmentations for test patients with ductal carcinoma in situ. 199

E.1. Predicted tumour contours for two invasive lobular carcinomas from the BIMAP dataset. 205

E.2. Predicted tumour contours for two ductal carcinoma in situ specimens from the BIMAP dataset. 206

E.3. Predicted tumour contours for two invasive carcinomas of no special type from the BIMAP dataset, pretreated with neoadjuvant chemotherapy. 207

E.4. Contingency tables for margin status prediction in the BIMAP dataset, shown separately for invasive lobular carcinoma (ILC), for ductal carcinoma in situ (DCIS), and for invasive carcinoma of no special type pretreated with neoadjuvant chemotherapy (NAC). 208

E.5. P-values (p) resulting from McNemar’s test to compare different approaches to predict the margin status for the BIMAP dataset, shown separately for invasive lobular carcinoma (ILC), for ductal carcinoma in situ (DCIS), and for invasive carcinoma of no special type pretreated with neoadjuvant chemotherapy (NAC). 209

E.6. Predicted tumour contours for two invasive lobular carcinomas from the BrIMA dataset. 212

E.7. Predicted tumour contours for two ductal carcinoma in situ specimens from the BrIMA dataset. 213

E.8. Predicted tumour contours for three neoadjuvant-treated (NAT) lumpectomy specimens from the BrIMA dataset. 214

E.9.	Contingency tables for margin status prediction in the BrIMA dataset, shown separately for invasive lobular carcinoma (ILC), for ductal carcinoma in situ (DCIS), and for all neoadjuvant-treated (NAT) cases combined.	215
E.10.	P-values (p) resulting from McNemar's test to compare different approaches to predict the margin status for the BrIMA dataset, shown separately for invasive lobular carcinoma (ILC), for ductal carcinoma in situ (DCIS), and for all neoadjuvant-treated (NAT) cases combined.	216
F.1.	Contingency tables for margin status prediction in the BIMAP dataset, using the leave-one-out strategy to select the operating point, separately for invasive carcinomas of no special type (NST), and for all breast cancer (BCa) subtypes combined.	219
F.2.	P-values (p) resulting from McNemar's test to compare different approaches to predict the margin status for the BIMAP dataset, using the leave-one-out strategy to select the operating point, shown separately for invasive carcinomas of no special type (NST), and for all breast cancer (BCa) subtypes combined.	220
F.3.	Contingency tables for margin status prediction in the BrIMA dataset, using the leave-one-out strategy to select the operating point, shown separately for invasive carcinomas of no special type (NST), and for all breast cancer (BCa) subtypes combined.	222
F.4.	P-values (p) resulting from McNemar's test to compare different approaches to predict the margin status for the BrIMA dataset, using the leave-one-out strategy to select the operating point, shown separately for invasive carcinomas of no special type (NST), and for all breast cancer (BCa) subtypes combined.	223

List of Tables

2.1. Diagnostic accuracy of standard-of-care intraoperative margin assessment techniques and emerging technologies according to meta-analyses [71–76]. 30

2.2. Overview of specimens included in the BIMAP and BrIMA studies per breast cancer (BCa) subtype, as well as the performance of micro-PET-CT for margin assessment across subtypes. 46

4.1. Registration components for the first three steps of the co-registration algorithm. 86

4.2. Quantitative metrics expressing the image alignment after the different co-registration steps. 97

4.3. Standardised uptake values (SUVs) of different breast tissue types as measured in [¹⁸F]FDG micro-PET images. 98

4.4. Ratios of standardised uptake values (SUVs) of different breast tissue types relative to healthy tissue, as measured in [¹⁸F]FDG micro-PET images. 100

5.1. Overview of all 150 Residual U-Nets (ResU-Nets) and 30 intensity thresholds (Intens thresh) trained in this work, with their respective input images, and the breast cancer (BCa) type they were trained to segment. 115

5.2. Metric values to evaluate the agreement between true and predicted tumour segmentations for test patients with invasive carcinoma of no special type (NST). 117

5.3. P-values (p) resulting from the Wilcoxon signed-rank test, comparing the performance of the ResU-Nets and intensity thresholding (Intens thresh) models across different model inputs for segmentation of invasive carcinoma of no special type (NST). 117

6.1. Metric values to evaluate the agreement between the true histopathological and predicted margin status for the BIMAP dataset, shown separately for invasive carcinomas of no special type (NST), and for all breast cancer (BCa) subtypes combined. 135

6.2. Metric values to evaluate the agreement between the true histopathological and predicted margin status for the BrIMA dataset, shown separately for invasive carcinomas of no special type (NST), and for all breast cancer (BCa) subtypes combined. 139

- 7.1. Calibration metrics for pixel-level tumour probabilities before and after Venn-ABERS calibration. 160

- 7.2. Metrics expressing the tumour segmentation and margin assessment performance of the Residual U-Nets (ResU-Nets) before and after Venn-ABERS calibration. 162

- B.1. Quantitative metrics expressing final alignment between the micro-CT images and histopathology whole slide images after co-registration with and without guidance of the micro-CT of the fixated lamellas. 189

- C.1. Overview of hyperparameters of the Residual U-Nets (ResU-Nets) obtained through grid search. 193

- D.1. Metric values to evaluate the agreement between true and predicted tumour segmentations for test patients with invasive lobular carcinoma (ILC) and ductal carcinoma in situ (DCIS). 200

- D.2. P-values (p) resulting from the Wilcoxon signed-rank test, comparing the performance of the ResU-Nets and intensity thresholding (Intens thresh) models across different model inputs for segmentation of invasive lobular carcinoma (ILC) and ductal carcinoma in situ (DCIS). 201

- E.1. Metric values to evaluate the agreement between the true histopathological and predicted margin status for the BIMAP dataset, shown separately for invasive lobular carcinoma (ILC), for ductal carcinoma in situ (DCIS), and for invasive carcinoma of no special type pretreated with neoadjuvant chemotherapy (NAC). 204

- E.2. Metric values to evaluate the agreement between the true histopathological and predicted margin status for the BrIMA dataset, shown separately for invasive lobular carcinoma (ILC), for ductal carcinoma in situ (DCIS), and for all neoadjuvant-treated (NAT) cases combined. 211

- F.1. Metric values to evaluate the agreement between the true histopathological and predicted margin status for the BIMAP dataset, using the leave-one-out strategy to select the operating point, shown separately for invasive carcinomas of no special type (NST), and for all breast cancer (BCa) subtypes combined. 219

F.2. Metric values to evaluate the agreement between the true histopathological and predicted margin status for the BrIMA dataset, using the leave-one-out strategy to select the operating point, shown separately for invasive carcinomas of no special type (NST), and for all breast cancer (BCa) subtypes combined. 222

1 | Introduction

1.1. Clinical rationale

Breast cancer (BCa) is the most commonly diagnosed cancer and the leading cause of cancer mortality in women worldwide [1–3]. While mastectomy used to be standard, most patients with early-stage BCa are now treated with breast-conserving surgery (BCS) followed by radiotherapy [4–8]. In BCS, or lumpectomy, only the tumour and surrounding tissue are removed. For early-stage disease, successful BCS provides overall and disease-free survival comparable to mastectomy, while offering improved cosmetic results and quality of life [4–7, 9–12]. Successful BCS requires complete tumour excision with a margin of surrounding healthy tissue, i.e. negative resection margins, to assure no cancer cells remain within the breast [7, 13, 14]. Positive margins, indicating residual tumour cells at the specimen border, remain the main risk factor for local recurrence after BCS [9, 15–19].

Sectional histopathology is the gold standard for margin assessment [20]. The specimen surface is marked with ink and according to the 2014 Society of Surgical Oncology and American Society for Radiation Oncology (SSO-ASTRO) consensus, “ink on tumour” defines a positive margin for invasive BCa [18]. Positive margins are found in approximately 20% of patients with invasive BCa after initial BCS, with reported rates ranging from 7% to 36% [7, 21, 22]. Because histopathological margin assessment typically requires several days, re-excision or mastectomy is often needed in case of positive margins to clear the breast of remaining tumour cells. These additional surgeries increase patient anxiety, morbidity, and healthcare costs [23–26].

To avoid secondary surgery, techniques for fast intraoperative margin assessment (IMA) are crucial to enable the removal of additional tissue during the same surgery if margins are suspected to be positive [27]. Several techniques have been explored for this purpose, including frozen sections, imprint cytology, intraoperative ultrasound, and specimen radiography [28, 29]. However, existing standard-of-care approaches have important limitations: they are either time-consuming, insufficiently accurate, costly, labour-intensive, or disruptive to the surgical workflow [28]. Recently, specimen imaging with high-resolution (micro-) positron emission tomography (PET), combined with computed tomography (CT), has emerged as a promising technique for IMA during BCS [30–33]. For this technique, the radiotracer [^{18}F]fluorodeoxyglucose ([^{18}F]FDG) is preoperatively administered to the patient. Afterwards, during BCS, a dedicated compact and mobile micro-PET-CT specimen scanner

is used to visualise the resected lumpectomy specimen in 3D inside the operating theatre, enabling rapid assessment of all margins.

The recent multicentre prospective BrIMA trial evaluated physicians' ability to intraoperatively interpret micro-PET-CT images of lumpectomy specimens and assess their margins [30, 32, 33]. Surgeons assessed 152 specimens of different BCa subtypes, achieving a sensitivity of 77% and specificity of 67% for positive margin detection [33]. The success rate of intraoperatively addressing positive margins with micro-PET-CT specimen imaging was significantly higher than with standard-of-care IMA techniques [33]. The average interpretation time was approximately one minute per 3D image [34]. Inter-reader reproducibility analysis showed high consistency when combining assessments from four surgeons, although variability remained in single-reader evaluations [33].

1.2. Research objectives

Although study results are promising, we hypothesise that a decision-support system to assist physicians in interpreting micro-PET-CT images could further improve inter-reader reliability, reduce intraoperative interpretation time, and potentially enhance sensitivity and specificity. The need for automated, artificial intelligence-based tools to support micro-PET-CT interpretation, has also been expressed by Lambert et al. [31] and Moraitis et al. [35]. In this work, we take initial steps towards such a clinical decision-support system by developing a deep learning (DL) model that automatically segments (pre)malignant breast tissue in micro-PET-CT images of lumpectomy specimens, thereby also enabling automated margin status prediction. Realising and evaluating such a DL model raises several research questions. The research contributions in this dissertation address the following four questions:

1. Can micro-PET-CT images of BCa specimens be co-registered with histopathology whole slide images (WSIs) precisely enough to (i) characterise [^{18}F]FDG uptake per breast tissue type and (ii) derive ground-truth tumour segmentations from histopathology?
2. Can a DL model trained on histopathology-derived ground-truth segmentations delineate (pre)malignant breast tissue in micro-PET-CT images more accurately than conventional PET intensity thresholding? And does combining micro-PET with micro-CT improve delineation over either modality alone?
3. Can automated margin status prediction from DL-based tumour segmentations match physicians' interpretations and outperform PET intensity thresholding, on a multicentre dataset spanning diverse BCa subtypes?

4. Can the tumour probability maps predicted by the DL model be calibrated so that the probabilities reflect the true likelihood of malignancy, yielding reliable spatially-resolved uncertainty estimates, without degrading segmentation or margin assessment performance?

1.3. Outline

This section outlines the structure of this dissertation. **Part I** of this thesis introduces the relevant clinical and technical background. **Chapter 2** elaborates on the clinical context of this work. It introduces breast anatomy and pathology, followed by BCa subtypes, epidemiology, clinical presentation, and diagnosis. Subsequently, BCa treatment strategies are discussed, with emphasis on BCS and the importance of achieving negative resection margins. The chapter concludes with an overview of current and emerging IMA techniques, focusing on micro-PET-CT specimen imaging, including the fundamental principles of PET and CT and the current clinical evidence supporting the use of this IMA technique in BCa. **Chapter 3** presents the technical background for developing DL models to analyse micro-PET-CT images. It introduces key concepts of DL, including deep artificial neural networks (ANNs), their supervised training process, and regularisation techniques. Application of DL in medical computer vision is then discussed, focusing on classification and semantic segmentation with convolutional neural networks (CNNs) such as ResNet and U-Net. Next, challenges associated with DL in medical image analysis are addressed, along with emerging trends in the field. Finally, this chapter briefly introduces classical iterative image co-registration algorithms used in this dissertation to generate ground-truth tumour segmentations.

Part II of this dissertation discusses obtained research results. **Chapter 4** presents a method to co-register micro-PET-CT specimen images with corresponding histopathology WSIs. This method is applied to BCa specimens from patients preoperatively injected with [^{18}F]FDG. The aligned data provides important insights into [^{18}F]FDG uptake across breast tissue types at high resolution. Furthermore, tumour annotations from histopathology can be transferred to the aligned micro-PET-CT images to obtain ground-truth breast tumour segmentations. This dataset is used in **Chapter 5** to develop a DL model for automated delineation of breast tumours in micro-PET-CT images. The performance of this DL model is compared with conventional PET intensity thresholding. This DL model is further validated for margin status predictions of lumpectomy specimens in **Chapter 6**. The DL model is used to segment breast tumours in previously unseen micro-PET-CT datasets, after which the margin status is determined based on whether the segmented tumour touches the specimen border. These predictions are compared with the

histopathological ground-truth margin status, as well as with physicians' assessments of the same images. **Chapter 7** investigates how the tumour probability maps predicted by the DL model can be calibrated so that the predicted probabilities reflect the true likelihood of malignancy, yielding reliable spatially-resolved uncertainty estimates to support risk-informed clinical decision-making.

Finally, **Part III** of this dissertation, consisting of **Chapter 8**, presents the general research conclusions and discusses future perspectives.

Part I

Clinical and Technical Background

2 | Micro-PET-CT specimen imaging for intraoperative margin assessment in breast-conserving surgery

2.1. Introduction

This chapter elaborates on the clinical background underlying the rationale of this work. It starts with an overview of normal breast anatomy, followed by an introduction to breast pathology, ranging from benign conditions to (pre)malignant disease. Subsequently, different histological and molecular subtypes of breast cancer (BCa), its epidemiology, and most common clinical symptoms are discussed. The diagnostic pathway of BCa is then explained, focusing on the so-called “triple-test”, which combines physical examination, medical imaging, and histopathological assessment of biopsies.

Next, an overview of local, locoregional, and systemic treatment strategies for BCa is provided. In the context of this PhD dissertation, particular emphasis is placed on the local treatment of BCa through breast-conserving surgery (BCS), which is the most common primary treatment for early-stage BCa. The importance of negative resection margins in successful BCS is discussed, together with the prevalence and clinical consequences of positive margins.

Subsequently, standard-of-care (SOC) and emerging techniques for intraoperative margin assessment (IMA) are reviewed, including their diagnostic performance in detecting positive margins, important advantages, and limitations. Finally, this chapter focuses on micro-PET-CT specimen imaging as a promising IMA technique. First, the fundamental principles of positron emission tomography (PET) and computed tomography (CT) are described. Afterwards, we explain how these modalities can be used to image surgically excised BCa specimens and we review the current clinical evidence supporting the use of micro-PET-CT specimen imaging as IMA technique in BCa.

2.2. Breast cancer

2.2.1. Breast anatomy

The breasts are located on the anterior thoracic wall and are present in both males and females, but get more developed in females during puberty [36]. They contain the mammary glands, essential for lactation [37]. The mammary glands are embedded in adipose (fatty) tissue and supported by connective (fibrous) tissue. This connective tissue starts from the pectoral fascia (a sheet of connective tissue covering the pectoral muscles), extends throughout the breast to provide structural support and shape, and terminates in the skin that surrounds the breast tissue. The mammary glands mainly consist of lobules that contain the milk-producing units and ducts that transport milk and terminate in the nipple. The nipple, which is primarily composed of smooth muscle fibres, is surrounded by a pigmented skin area known as the areola [36]. The areola contains sebaceous glands that secrete an oily substance to protect the nipple during breastfeeding. Figure 2.1 illustrates the most important anatomical structures of the breast.

When studying the mammary glands in more detail, the breast can be subdivided in 15 to 20 lobes [37]. Each lobe is connected to one lactiferous duct that terminates at the nipple. Deeper in the breast, lactiferous ducts branch into smaller terminal ducts. Surrounding these terminal ducts, lobules are found. There can be up to 40 lobules per lobe [37]. A lobule together with its associated terminal duct forms the terminal duct lobular unit (TDLU), which represents the functional unit of the breast [37, 38]. Lobules consist of smaller glandular units, called acini. Breast acini are lined by a double epithelial layer: an inner layer of milk-producing luminal epithelial cells, and an outer layer of contractile myoepithelial cells [38]. These epithelial structures are enclosed by a basement membrane, separating them from the surrounding tissue. While ducts have a similar structure, the inner layer of epithelium shifts from a secretory to a transport function as duct size increases. The relation between these anatomical structures is visualised in Figure 2.1. As discussed later on, the TDLUs play an important role in BCa, as BCa most commonly starts in the epithelial cells lining the TDLUs [37].

Furthermore, the breast contains an extensive lymphatic network that drains lymphatic fluid towards multiple groups of lymph nodes located within and surrounding the breast tissue [37]. Of particular clinical importance are the axillary lymph nodes (indicated in Figure 2.1), which are located in the axilla (armpit) and represent the primary drainage pathway of the breast [1, 37]. The lymphatic network can play an important role in the metastatic spread of tumour cells to other parts of the body. Finally, networks of arteries and veins, as well as neurovascular bundles also run through the breast.

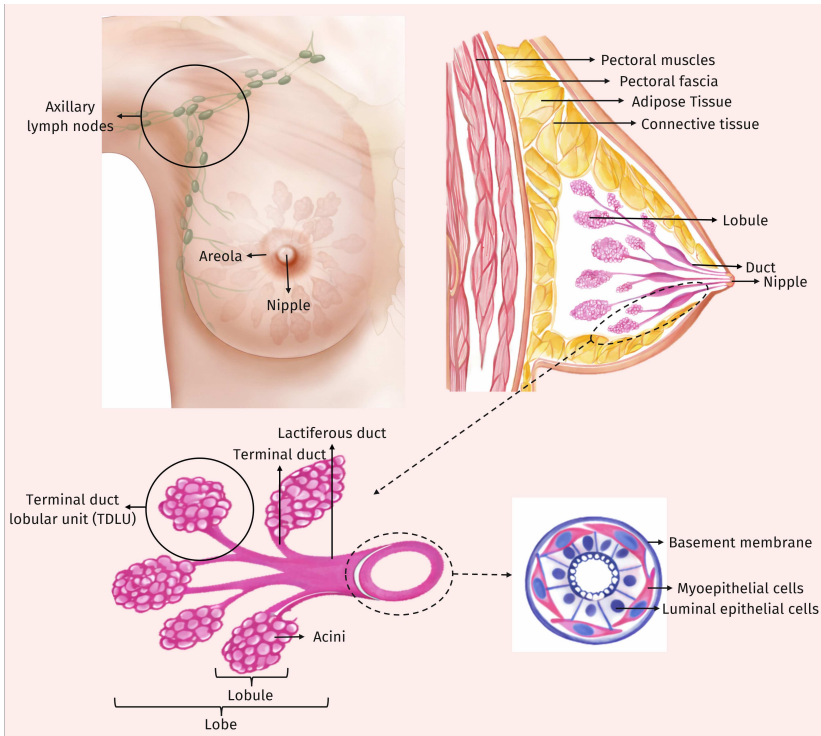


Figure 2.1. Anatomy of the female breast. Figures adapted from [38, 39].

2.2.2. Breast pathology

A wide variety of breast pathologies exist. In this section we discuss only the most common benign, premalignant, and malignant conditions. While premalignant and malignant conditions can occur in any type of breast tissue, we focus on the most common (pre)malignancies which originate from the epithelial cells lining the ducts and lobules of the breast.

2.2.2.1. Benign conditions

Approximately 50% of women will develop a benign breast lesion in their lifetime, and up to 90% of palpable breast masses in menstruating women are benign [1]. Common benign breast lesions are fibroadenomas and cysts. Fibroadenomas are lesions formed through excessive growth of epithelial and connective breast tissue [1, 37]. They present as solid, clearly defined, and mobile nodules. On medical images they look round or oval and homogeneous. In women younger than 21, 60% of nodules are fibroadenomas [37]. Less than 10% of these nodules grow, and one out of three lesions disappears over time [37]. Fibroadenomas do not

require treatment, but can be further examined or resected if they grow or are too large. Cysts are cavities filled with fluid, frequently occurring in women aged 35-55 years [37]. They can cause tension and pain. Cysts are well-defined and mobile structures. They appear on echography as clearly defined dark nodules. The fluid can be evacuated through needle aspiration.

2.2.2.2. Premalignant conditions

Premalignant or high-risk epithelial breast lesions are benign epithelial proliferations associated with an increased risk of subsequent breast carcinoma [37]. When epithelial cells proliferate within the milk ducts, this can be classified as atypical ductal hyperplasia or ductal carcinoma in situ (DCIS). The distinction between both is primarily based on the extent and architectural distribution of the lesion [37, 40]. Atypical ductal hyperplasia is associated with an approximately fourfold increased risk of BCa [37]. When epithelial cells proliferate within the lobules, this can be classified as atypical lobular hyperplasia or lobular carcinoma in situ (LCIS). When more than 50% of the acini within a lobule are dilated, the lesion is classified as LCIS [37]. Other high-risk lesions include papillary lesions and radial scars or complex sclerosing lesions [37]. In the interest of this work, DCIS and LCIS are discussed in more detail below.

Ductal carcinoma in situ (DCIS) manifests as a proliferation of malignant epithelial cells, confined within the lumen of the milk ducts without disruption of the basement membrane [37]. Although DCIS is a non-invasive carcinoma, it represents a precursor lesion with unpredictable potential for progression to invasive BCa. DCIS is most frequently detected through screening as it is often asymptomatic. Small bright spots are commonly seen on mammography. They are microcalcifications, dense deposits of calcium often caused by blocked milk ducts [41]. Treatment of DCIS is similar to the treatment of invasive BCa, often entailing surgical resection of the lesion with a wide margin, followed by radiotherapy.

Lobular carcinoma in situ (LCIS) is characterized by a proliferation of atypical epithelial cells, confined within the lobules with an intact basement membrane [37]. LCIS is not regarded as a precursor of invasive carcinoma, but is considered a marker of increased risk for later development of invasive carcinoma [4]. Surgical excision or treatment is often not required.

2.2.2.3. Malignant conditions

The majority of epithelial breast malignancies are invasive carcinomas [37]. These tumours are characterized by invasion beyond the basement membrane surrounding the lobules and ducts. The most frequent invasive subtypes are invasive carcinoma of no special type (NST), and invasive lobular carcinoma (ILC), which are discussed in more detail below [37]. Less common special histological subtypes include tubular, mucinous, and medullary carcinoma [37].

Invasive carcinoma of no special type (NST) typically arises from the ductal epithelium and is characterised by tumour cell invasion beyond the basement membrane [37]. It commonly presents as tumour cells that clump together and form lumps. The diagnosis is established by exclusion, when the cancer cells lack any distinct histological features characteristic of special subtypes of invasive carcinoma. This subtype was formerly known as invasive ductal carcinoma (IDC).

Invasive lobular carcinoma (ILC) is the most common special type of invasive BCa [42]. Typically, these tumours originate in the lobules, and break through the basement membrane. ILC lesions often show a non-cohesive branching growth pattern, with tumour cells arranged in single-file lines [42]. These patterns can mimic connective and glandular tissue on medical images, making ILC hard to detect [37].

2.2.3. Breast cancer classification

BCa can be classified in different ways. Depending on the clinical question, different characteristics might be of importance to define the correct diagnosis, prognosis, and treatment strategy. A full classification includes the histological subtype, molecular subtype, grade, and stage of the tumour.

2.2.3.1. Histological classification

Histological classification of BCa is based on cell characteristics and architectural patterns seen when studying tissue biopsies under the microscope [4]. While BCas can have many different histologies, we focus on the majority group derived from the epithelial cells lining the ducts and lobules. At a simplified level, as described in the previous section, the most common types of BCa can be classified according to their site of origin (lobular or ductal), and the presence or absence of invasion beyond the basement membrane (invasive or in situ). As illustrated in Figure 2.2, the following histological BCa types can be distinguished:

- **Ductal carcinoma in situ (DCIS)** arises from the ductal epithelium and is characterised by preservation of the basement membrane.
- **Lobular carcinoma in situ (LCIS)** originates from the lobular epithelium and likewise leaves the basement membrane intact. However, LCIS is not considered true BCa, but is seen as a marker of increased risk for later BCa development [4].
- **Invasive carcinoma of no special type (NST)** arises from the ductal epithelium and invades through the basement membrane.
- **Invasive lobular carcinoma (ILC)** originates from the lobular epithelium and likewise invades through the basement membrane.

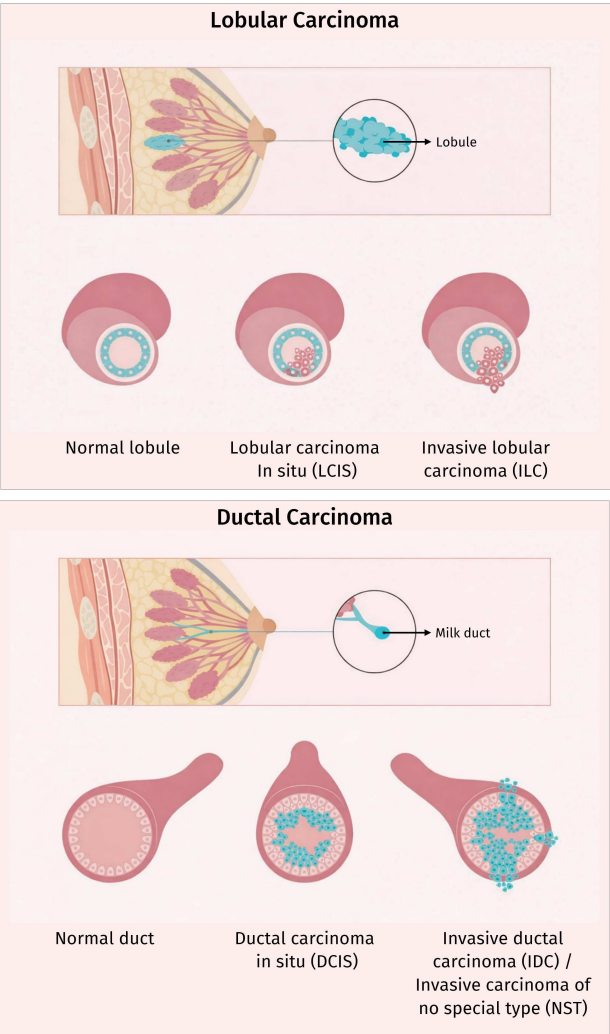


Figure 2.2. Histological breast cancer subtypes. Figure adapted from [43].

2.2.3.2. Molecular classification

The molecular classification of BCas is of high clinical relevance, as it can guide treatment selection. A first distinction is made between BCas with positive and negative hormone receptor (HR) status, indicating whether the tumour cells express oestrogen receptors (ERs) and/or progesterone receptors. The growth of HR-positive tumours is driven by oestrogen and/or progesterone. Second, BCas are also classified according to their human epidermal growth factor receptor 2 (HER2) status. HER2-positive tumours show an overexpression of the HER2 protein, typically associated with aggressive cancers. Third, tumour cell proliferation is commonly assessed by evaluating the expression of the proliferation marker Ki-67. Based on these molecular features, four molecular subtypes are distinguished:

- **Luminal A** BCa is positive for HR, while negative for HER2, in combination with low levels of Ki-67.
- **Luminal B** BCa is positive for HR, while negative for HER2, in combination with high levels of Ki-67.
- **Triple negative** BCa is negative for both HR and HER2.
- **HER2 enriched** BCa is positive for HER2, in combination with either positive or negative HR status.

Luminal A type showed the best short-term prognosis with the best five-year cancer-specific and overall survival, while triple negative BCa had the worst short-term survival [44–46].

2.2.3.3. Grading

How malignant and aggressive a breast tumour is, is often expressed using the histological grading system [47]. In case of invasive BCas (NST and ILC), the Nottingham histologic grading system is used [4]. The overall grade is defined as a sum of scores defined through semi-quantitative evaluation of three morphological features: tubule formation, nuclear pleomorphism, and mitotic count [4, 48]. Lower histological grade showed a strong correlation with better prognosis [48]. In general, a low-grade (grade 1), well differentiated tumour grows more slowly and is less likely to spread, while high-grade (grade 3), poorly differentiated tumours tend to grow and spread more quickly and may require immediate or more aggressive treatment [47]. In case of DCIS, a slightly modified grading system is used to score intraluminal necrosis, nuclear atypia, and mitotic activity [4]. Higher DCIS grades indicate increased potential for locoregional recurrence and for progression to invasive BCa [4].

2.2.3.4. Staging

The BCa stage defines the anatomic extent of cancer [37]. Staging is performed using the Tumour-Nodes-Metastasis or TNM-system [37]. The clinical (cTNM) stage is defined through medical imaging and blood tests. A more accurate final pathological (pTNM) stage can only be defined after resection and histopathological examination of the tumour, lymph nodes, and metastases. The different components of the TNM-stage are as follows:

- **T-stage** (Tis,T1-T4) reflects the maximal diameter of the primary tumour, often measured through echography or mammography [4]. In situ cancers are indicated with Tis [4]. Tumours with a T-stage of Tis or $\leq T_2$, corresponding to a tumour size ≤ 5 cm, are considered early-stage BCas [4].
- **N-stage** (N0-N3) reflects involvement of lymph nodes nearby the primary tumour site [4]. Axillary lymph nodes are often examined using echography.
- **M-stage** (M0-M1) indicates the presence of distant metastases, which most frequently occur in the bones, lungs, brain, or liver [4]. Whole-body PET-CT can be used to check for distant metastases in high-stage BCas [4].

2.2.4. Epidemiology

2.2.4.1. Incidence

BCa is the most commonly diagnosed cancer in women, and is the leading cause of cancer mortality in women worldwide [1–3]. In 2020, 2.26 million new BCas were diagnosed worldwide, and 685 000 women died from BCa [2]. Almost two-thirds of those deaths were recorded in less-developed regions. In high-income countries, overall 5-year survival from BCa is well over 80%, while it drops to around 60% for middle-income, and 40% for low-income countries [2, 4].

For Belgium in particular, yearly around 10 000 new cases of BCa are diagnosed [37]. BCa is the most important cause of death for women aged 40 to 55 years [37]. A study by van Walle et al. [45] showed that for all new BCas diagnosed in 2014 in Belgium, the mean age of patients was 62, with 1/3 patients being 70 years or older, and 1/5 patients being younger than 50. The unadjusted five-year relative survival proportion for the Belgian cohort was 91.4% [45]. For Belgian women, 1/10 develop BCa in their lifetime, while the risk of dying from BCa is around 3.2% [37]. These numbers are comparable to the numbers for the US, where the American Cancer Society estimated that 1/8 women develop BCa in their lifetime, and the risk of dying from BCa is 3.4% [1, 4]. While BCa can occur in men, it is rare. Less than 1% of all BCas is diagnosed in men [4, 49].

2.2.4.2. Risk factors

Several factors increase the risk of BCa [2, 4, 37]:

- Increasing age (incidence increases between the age of 30 and 65)
- Geographic location (higher incidence in North America and Western Europe, lower incidence in East Asia)
- Early menarche and late menopause
- Nulliparity and advanced age at first full-term pregnancy
- Previous history of benign breast conditions
- Exposure of the breast to ionizing radiation at a young age
- Alcohol consumption
- Postmenopausal obesity
- Long-term use of hormonal (contraceptive) medication
- Genetic predisposition (e.g. BRCA1 or BRCA2 mutations)

2.2.4.3. Prevalence of breast cancer subtypes

The following ranges in the prevalence of histological BCa subtypes were found when combining results from study cohorts in Belgium [37, 45], the US [4, 40], the UK [50], and Western India [51]:

- DCIS: 15%-20%
- Invasive carcinomas: 80%-85%
 - NST: 73%-96%
 - ILC: 10%-15%
 - mixed NST + ILC: 5%

For the prevalence of molecular subtypes, we found the following ranges when combining results from study cohorts in Australia [44], Belgium [45], the US [52], West India [51], and China [46]:

- Luminal A: 37%-54%
- Luminal B: 7.6%-16.7%
- Triple negative: 10%-26%
- HER2 enriched: 11%-15%

The stage of BCa at diagnosis varies substantially worldwide, and is generally more advanced in low-income countries, mainly due to limited access to regular screening. In Belgium, regular screening attendance was associated with a sixfold lower risk of late-stage BCa diagnosis compared with irregular attendance or non-attendance [53]. A recent global meta-analysis reported early-stage invasive BCa diagnosis ranging from 14%-46% in Sub-Saharan Africa to 82%-90% in North America, with intermediate values in Central and South America (52%-65%), Central and West Asia (57%-59%), East and South Asia (59%-76%), North and

West Europe (76%-86%), East and South Europe (61%-83%), and Oceania (82%) [54]. Similarly, distant metastatic disease was more common in Sub-Saharan Africa, Central and South America, and Central and West Asia (10%-30%), and less frequent in North America, Europe, and Oceania (<10% in most countries) [54]. Data on the distribution of BCa grades at diagnosis are limited. However, younger patients (<40 years old) are reported to have a higher risk of presenting with high-grade BCa [55].

2.2.5. Symptoms

Most common symptoms of BCa include [4, 37]:

- Palpable mass in the breast
- Nipple discharge
- Skin changes, retraction of the skin or nipple, or oedema of the skin
- Pain, although most patients are pain-free at first diagnosis
- Axillary adenopathy or metastases causing illness, although rare at first presentation

2.2.6. Diagnosis

When a patient presents with symptoms of BCa, or when an anomaly was detected during screening, the patient needs to be further examined to confirm the diagnosis. The evaluation involves the “triple test”, consisting of physical examination, imaging, and histological assessment [37]. When there are discrepancies between these three pillars of BCa diagnosis, further examinations should be performed.

2.2.6.1. Physical examination

Palpation of the breast, as well as the axillary and clavicular lymph nodes, is performed to detect masses and signs of infection or swelling [37]. In addition, the breast skin is inspected for redness, retraction of the skin or nipple, asymmetry, oedema, nipple eczema, and nipple discharge [37].

2.2.6.2. Medical imaging

When physical assessment showed potential malignancies, further examination is performed using medical imaging techniques. All medical imaging modalities assess the presence of masses, as well as asymmetries and architectural distortions within the breast tissue [4]. Malignant lesions typically appear as dense masses with poorly delineated and irregular shapes [4]. The most commonly used imaging techniques in BCa diagnosis are mammography, ultrasound (US), and magnetic resonance imaging (MRI), with PET-CT imaging reserved for selected cases [37].

Mammography involves planar 2D radiography imaging of the breast. A small dose of X-rays is sent through the breast, and tissues with different densities will lead to different levels of X-ray attenuation. Both lesions and microcalcifications appear as bright high-density structures on mammograms. Breast tomosynthesis is a related technique where emulated 3D images are reconstructed from several 2D projection images taken at different angles. Diagnostic or clinical mammography is performed after detection of a breast abnormality, whereas screening mammography is performed in asymptomatic women to detect subclinical disease. In Flanders, screening mammography is offered every two years to women aged 50-69 years, and has been associated with an approximate 20% reduction in BCa mortality [37]. An important limitation of mammography is its reduced performance in women with dense breasts [4]. False positive findings occur in approximately 10-15% of cases, with even higher numbers for dense breasts [37]. On the other hand, dense glandular tissue can also obscure small lesions, leading to false negative results. As such, mammography is less suited in younger women who often present with dense breasts. Furthermore, although the radiation dose is low, mammography should be avoided in women under the age of 40 [37].

Ultrasound (US) or echography generates images from echoes of high-frequency sound waves reflected by tissues with different acoustic properties. On US, tumours often present as hypoechoic or isoechoic masses [4]. In addition, tumours often demonstrate hypervascularity and increased stiffness, which can be assessed using specialised US techniques like Doppler and elastography [4]. US is commonly performed when mammography findings were inconclusive, and is the preferred first-line imaging modality in women under the age of 40, as well as in pregnant or lactating women [37]. In addition, US can also be used for screening in women with dense breasts. US demonstrates higher sensitivity for the detection of NST BCa in dense breasts than mammography, although it remains less sensitive compared to MRI [4]. Finally, US is also routinely used to guide biopsy and clip or wire placement [4].

Magnetic resonance imaging (MRI) uses a strong magnetic field and radiofrequency waves to generate detailed soft tissue images from signals emitted by relaxing hydrogen nuclei. Tumour appearance depends on the MRI sequence, with lesions appearing isointense on T1-weighted images and iso- to hypointense on T2-weighted images with fat suppression, while necrosis or oedema typically shows T2 hyperintensity [4]. Functional MRI techniques can further assess tumour characteristics such as cellular density. While MRI is a very sensitive modality to detect breast lesions, it is time-consuming, expensive, and prone to false positive results [37]. MRI can be used in case of discordance between mammography and US results, to differentiate recurrence from postoperative fibrosis, to assess response to neoadjuvant chemotherapy (NAC), to evaluate multifocality, and for screening in high-risk patients [4, 37].

PET-CT imaging with [^{18}F]fluorodeoxyglucose ([^{18}F]FDG), a radioactive glucose analogue, combines detection of metabolically active lesions (PET) with detailed anatomical localisation (CT). Its role in BCa diagnosis is limited, as it is not indicated in early-stage BCa due to high false positive and false negative rates [37]. False positive results may be caused by [^{18}F]FDG uptake in regions with inflammation. False negative results may arise from the limited spatial resolution of whole-body PET scanners (3-8 mm), which can result in small lesions (≤ 1 cm) being missed, or their uptake being underestimated due to partial volume effects [4]. Uptake of [^{18}F]FDG in breast tumours varies widely depending on the BCa subtype. Uptake is generally higher in NST tumours than in DCIS and ILC [4]. In NST tumours, high [^{18}F]FDG uptake has been associated with HER2 positive, HR negative, triple negative, high Ki-67, and high-grade tumours [56]. PET-CT is mainly used for restaging patients with known or suspected recurrence and for staging of advanced metastatic disease [37]. Other molecular breast imaging techniques exist but have a limited role in BCa diagnosis and screening [4].

BI-RADS (Breast Imaging Reporting and Data System) is a standardized reporting system used by radiologists to score breast imaging findings [37]:

- BI-RADS 0: inconclusive
- BI-RADS 1: normal
- BI-RADS 2: benign changes
- BI-RADS 3: most likely benign, follow-up required
- BI-RADS 4: suspicious for malignancy, biopsy required
- BI-RADS 5: most likely malignant

2.2.6.3. Biopsy and histopathological assessment

In cases where breast imaging shows a solid mass in women over 35 years, biopsy is performed to obtain a definitive diagnosis [37]. In younger women, biopsy may also be recommended in case of increased BCa risk or diagnostic uncertainty [37]. A small sample is taken from the tumour, typically under US guidance, and studied under the microscope to assess the presence of tumour cells. Fine needle aspiration is a technique where a thin hollow needle is used to collect cells from a suspicious lesion or lymph node. While it allows for fast diagnosis and is minimally invasive, the small sample of single cells does not allow for a full histological evaluation. Core needle biopsy involves obtaining tissue cores using a wider, spring-loaded hollow needle, allowing for accurate diagnosis and BCa subtype differentiation. Core needle biopsy is the standard technique prior to surgery and provides important information for surgical planning, staging, and assessment of the need for neoadjuvant therapy. Vacuum-assisted biopsy is a technique where a wide hollow needle with vacuum suction is used to aspirate larger volumes of suspicious tissue. It is often used in case of challenging lesions with microcalcifications. Mammography of the aspirated tissue can confirm the presence of microcalcifications to assure a representative tissue fragment was obtained.

2.2.7. Treatment

Once BCa is diagnosed, a treatment plan is defined, selecting a combination of local, locoregional, and/or systemic therapies based on the clinical presentation.

2.2.7.1. Local treatment: Surgery

The standard primary treatment for BCa is surgical resection of the breast tumour and staging of axillary lymph nodes [37]. Previously, mastectomy, involving removal of the entire breast, used to be the standard treatment. However, prospective research has shown that local excision of the tumour followed by radiotherapy can be equally effective, while offering superior cosmetic results [37]. This procedure, also called lumpectomy or breast-conserving surgery (BCS), is discussed in detail in Section 2.3, together with the importance of the resection margin status. For large or multifocal tumours, mastectomy may still be required. More than 25% of breast tumours are non-palpable or clinically occult [57]. For non-palpable tumours, additional guidance is required for intraoperative localisation. To direct tumour resection, a guide wire, magnetic seed, or radioactive seed

can be placed inside the tumour before surgery using US or mammography guidance [37].

As axillary lymph node status is an important prognostic factor and guides the need for adjuvant therapy, lymph nodes are also often resected during surgery. Typically, the sentinel node, the first draining axillary lymph node, is removed during surgery for staging purposes. A common technique to identify the sentinel node is to inject a ^{99}Tc -labelled radiotracer into the tumour, which emits gamma photons [58]. Once drained to the lymph nodes, a scintigraphy is performed to mark the location of the sentinel node on the skin. During surgery, a gamma probe is used to localise the sentinel node for resection. Complete axillary dissection is reserved for patients with suspected nodal invasion due to important side effects of this procedure like pain and lymphoedema [37].

2.2.7.2. Locoregional treatment: Radiotherapy

Radiotherapy is standard after BCS and may also be indicated after mastectomy [4]. The breast is irradiated with high-energy X-rays with the goal of eradicating subclinical residual disease and minimizing the risk of locoregional recurrence. In case of extensive disease, parts of the thorax and regional lymph nodes may also be irradiated [37]. It has been shown that radiotherapy after BCS halved the chance of locoregional recurrence after 10 years and reduced mortality after 15 years by 17% [37]. Common side effects include redness of the skin and radiation pneumonitis.

2.2.7.3. Systemic treatment

Systemic treatment refers to treatment with drugs that circulate via the bloodstream to act throughout the body. Patients with positive lymph nodes or large tumours (> 2 cm) often receive adjuvant systemic therapy after surgery to eradicate micro-metastases, reduce the risk of (locoregional) recurrence, and increase overall BCa survival [37]. Micro-metastases are detected in 2-30% of patients with negative lymph nodes, and in 50-90% of patients with positive lymph nodes [37]. Systemic therapy may consist of hormonal therapy, chemotherapy, targeted therapy, or a combination thereof.

Hormonal therapy slows down the growth of hormone-sensitive (oestrogen receptor (ER) and/or progesterone receptor positive) breast tumours by blocking their stimulation through sex hormones [37]. These treatments are highly effective, easy to administer, and well tolerated. In pre-menopausal women, the ovarian oestrogen production can be suppressed through medication or ovariectomy. Selective oestrogen receptor modulators (SERMs) are a class of drugs that can reduce tumour

growth in both pre- and postmenopausal women, but have important side effects as they increase the risk of thromboembolic events and endometrial cancer [37]. SERMs were shown to increase overall survival by 13%, and halve the risk of recurrence after 15 years [37]. Another type of hormone therapy drugs used to minimize the concentration of oestrogen in postmenopausal women are aromatase inhibitors, which block the aromatase enzyme that converts androgen to oestrogen [37]. While risk of recurrence after five years is lower in case of treatment with aromatase inhibitors compared to SERMs, this does not translate into an improvement in overall survival [37]. Common side effect include menopausal symptoms and an increased risk of osteoporosis [37].

Chemotherapy aims to inhibit tumour cell proliferation and induce tumour cell death by targeting rapidly dividing cells. While adjuvant chemotherapy is shown to halve the risk of recurrence, it has significant side effects, including fatigue, nausea, bone marrow suppression, increased risk of leukaemia, cardiovascular disease, premature menopause, and infertility [37]. Chemotherapy tends to be more effective in younger patients, who often present with aggressive tumours with low ER expression, and in patients with HER2 positive BCas or positive lymph nodes, who have an increased risk of metastases [37].

Targeted therapy acts on specific molecular features of tumour cells, thereby sparing normal cells. In HER2-enriched BCas, HER2 inhibitors can be used to slow down tumour growth and differentiation, reducing the risk of recurrence by approximately 50% [37].

The traditional order of therapy is to first perform surgical resection of the tumour, start chemotherapy when the wound has healed (typically after 4-6 weeks), and afterwards perform radiotherapy in combination with hormone therapy. However, in case of large tumours, neoadjuvant therapy, systemic therapy before surgery, may be required to shrink the tumour and make it operable. Neoadjuvant therapy often consists of neoadjuvant chemotherapy (NAC), but can also involve treatment with aromatase inhibitors in patients with large hormone-sensitive tumours for whom chemotherapy is contraindicated.

For patients with metastatic BCa, complete curation is generally not achievable, and palliative treatment is started aiming to prolong survival while preserving quality of life [37]. Surgery is typically not performed for these patients. Instead, systemic therapy is used as main treatment. Hormone therapy is used for hormone-sensitive tumours when metastatic disease is limited and not immediately life-threatening, whereas chemotherapy is indicated in other cases. In addition, radiotherapy can be employed to alleviate symptoms of metastatic lesions.

2.3. Breast-conserving surgery (BCS) and positive margins

As discussed in Section 2.2.7.1, breast-conserving surgery (BCS) is the primary local treatment for many BCa patients. Given the central role of BCS and its associated risk of positive resection margins in this work, these concepts are discussed in more detail in this section.

2.3.1. Principles of successful BCS

BCS, also called lumpectomy, involves local surgical excision of the breast tumour with maximal sparing of healthy breast tissue as illustrated in Figure 2.3. The majority (61%-75%) of patients with early-stage BCa is nowadays treated with BCS followed by radiotherapy [4–8]. For early-stage BCa, successful BCS is considered as effective and safe as mastectomy in terms of overall and disease-free survival, while BCS provides improved cosmetic outcomes and quality of life [4–7, 9–12]. The consensus is that successful BCS entails complete excision of the primary tumour with a margin of surrounding healthy tissue to assure no cancer cells remain within the breast [7, 13, 14]. Hence, the margin status, expressing adequacy of tumour excision, is the most important determinant of successful and safe BCS.

2.3.2. Margin status in BCS

The margin status of a breast tumour specimen is classified as positive or negative depending on whether tumour cells are present at the resection border. As illustrated in Figure 2.3, a positive margin indicates the presence of tumour cells at the specimen border, whereas a negative margin indicates a rim of healthy tissue at the border. Positive resection margins remain the most important risk factor for local recurrence after BCS [9, 15–19]. Vos et al. reported a 5-year locoregional recurrence risk of 2%-7% for negative margins compared with 0%-22% for positive margins [15, 18]. Similarly, Park et al. observed an 8-year locoregional recurrence risk of 7% for negative margins compared with 14%-27% for positive margins [59]. Therefore, accurate determination of the margin status has important diagnostic and prognostic value and relies on histopathological assessment of the excised specimen.

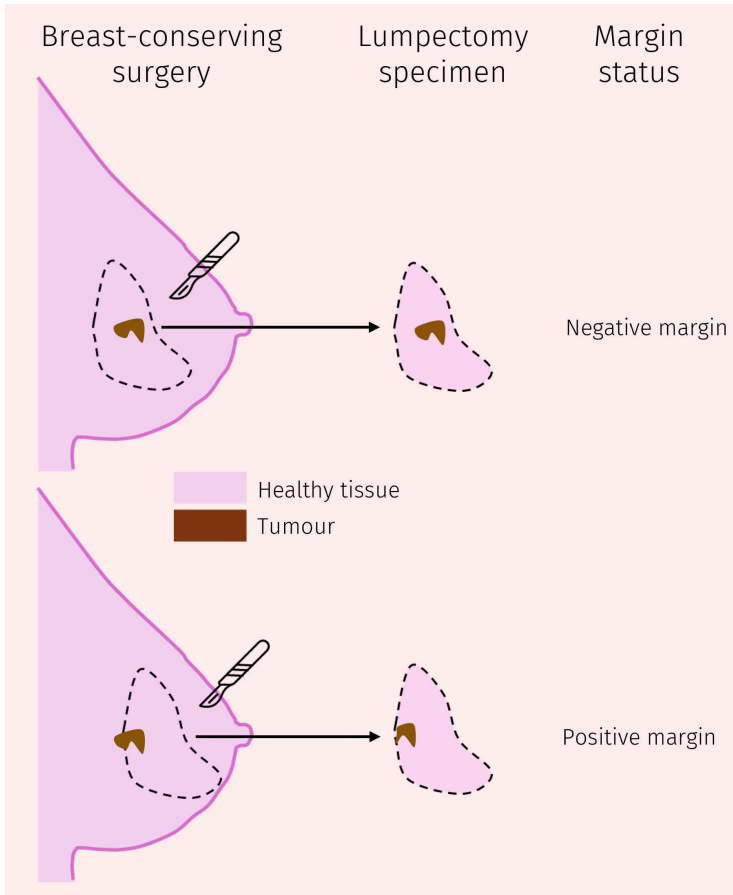


Figure 2.3. Illustration of breast-conserving surgery or lumpectomy and the associated surgical margin status.

2.3.3. Histopathology as gold standard for margin assessment

Sectional histopathology is considered the gold standard for margin assessment as it allows for detailed examination of the excised tissue at the cellular level [20]. The histopathological workflow is illustrated in Figure 2.4. Immediately after surgical excision, the breast specimen is transported to the histopathology department where the specimen undergoes several processing steps. First, the entire specimen surface is inked, using different colours to indicate the orientation of the specimen in the patient's body. Next, the specimen is sliced into sections of approximately 2 mm thickness, also called lamellas. Based on visual and tactile feedback, the pathologist selects suspicious lamellas for further analysis. These lamellas are then enclosed in tissue cassettes, meshed plastic boxes. If necessary, lamellas are further cut into halves or quarters

to fit the cassettes. After fixation in buffered formalin for at least 12 h, the tissue is embedded in paraffin. From each paraffin block, one thin tissue section of approximately 5 μm thickness is cut, mounted on a glass slide, and coloured using a haematoxylin and eosin (H&E) staining. Haematoxylin stains cell nuclei purple/blue, while eosin stains cytoplasm and the extracellular matrix red/pink. Finally, these glass slides are studied under a microscope at high magnification to differentiate between tumour and normal cells. This detailed examination allows precise measurement of the minimum distance between tumour cells and the inked resection margin. The histopathological workflow may vary slightly between hospitals. In some centres, the breast specimen is first fixated before being sliced into lamellas.

Current guidelines based on the 2014 Society of Surgical Oncology and American Society for Radiation Oncology (SSO-ASTRO) consensus recommend a tumour-free margin of more than 0 mm for invasive carcinoma (NST or ILC) and at least 2 mm for DCIS if postoperative radiotherapy is planned [18, 22, 60]. As such, the definition of a negative margin for invasive BCa is simply defined as “no ink on tumour” [18]. Complete histopathological assessment requires at least 16 hours and usually even 3-5 days, meaning positive margins are only detected long after surgery has been completed [20].

2.3.4. Limitations of histopathology

Despite being considered the gold standard, histopathology has inherent limitations that can lead to inaccurate margin status assessment in some cases. One important limitation is sampling bias: during histopathological examination, it is only feasible to analyse several tissue sections from the entire surgical specimen. As a result, approximately only 0.5% of the total specimen is examined microscopically [61]. This restricted sampling may result in false negative margin assessments.

Another challenge is specimen disorientation caused by miscommunication between surgeon and pathologist. When positive margins are found, it is crucial to retrieve the original position and orientation of the specimen in the patient's body, as this information can guide later re-excision of tumour cells remaining within the breast. However, studies have shown significant discordance in specimen orientation between surgeons and pathologists [62]. Traditionally, surgeons mark lumpectomy specimens with sutures to indicate orientation. Using suture-based orientation, Altman et al. reported that only in 20% of re-excisions after positive margin identification, residual disease was found [63].

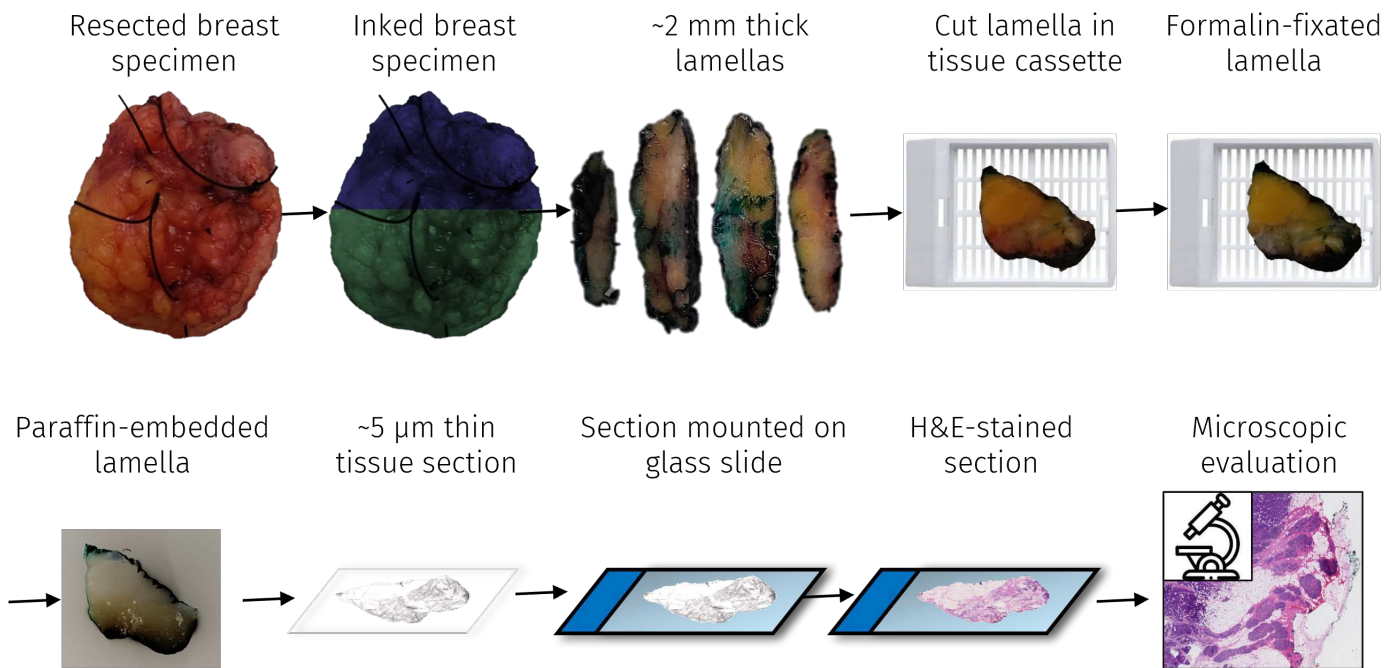


Figure 2.4. Histopathological processing of a breast cancer specimen.

2.3.5. Prevalence and impact of positive margins

Despite advances in surgical techniques, positive margins after BCS remain common. Reported positive margin rates after histopathological assessment of breast specimens vary widely in literature. Following initial BCS, positive margin rates for invasive BCa range from 7% to 36% [7, 21, 22], and Wilke et al. observed positive margins in approximately 24% of patients with early-stage BCa [64]. In general, higher rates are reported for ILC and DCIS, due to their diffuse and infiltrative growth patterns [64]. Positive margin rates of 18%-60% have been reported for ILC [65–67], and rates of 4%-52% have been reported for DCIS [21, 68–70]. Several other factors have been associated with an increased risk of positive margins, including younger patient age, larger tumour size, multifocal disease, and negative ER status [7, 64, 66, 68, 69]. The influence of tumour grade on the risk of positive margins is unclear, and contradicting results are reported in literature [68, 69].

Because of the increased risk of locoregional recurrence and the postoperative nature of histopathological margin assessment, 77%-99% of patients with positive margins have to undergo re-surgery to remove remaining tumour cells from the breast [21]. A common technique is cavity shaving, in which a thin layer of tissue is excised from the wall of the surgical cavity at the site where a positive margin was identified during histopathological evaluation. While re-excision does not result in poorer outcome in terms of survival, it can increase patient anxiety and morbidity. Furthermore, re-surgery is associated with a delay in the administration of adjuvant therapy, an increased risk of surgical complications and wound infections, poorer cosmetic outcomes, and increased healthcare costs [23–26].

2.4. Intraoperative margin assessment (IMA)

To avoid re-surgery, techniques for fast intraoperative margin assessment (IMA) are crucial, such that additional tissue can immediately be removed during the same surgery in case of suspected positive margins [27]. In this section we discuss several IMA techniques used in routine clinical practice, as well as other emerging imaging-based techniques. An ideal IMA technique should fulfil the following criteria:

- **High diagnostic accuracy:** Reliable discrimination between benign and malignant tissue is essential, as false positives may lead to unnecessary removal of healthy tissue causing impaired cosmetics and increased morbidity, whereas false negatives result in positive margins and the need for re-surgery.
- **Full 3D specimen evaluation with sufficient spatial resolution:** The entire surgical specimen should be evaluated to avoid sampling bias

and to determine both margin status and margin width, preferably at millimetre resolution as required for accurate IMA in DCIS.

- **Rapid intraoperative assessment:** Results should be rapidly available to minimise delay in surgery.
- **Workflow compatibility and ease of implementation:** The technique should easily integrate into the surgical workflow. Operator dependence and operator training requirements should be kept to a minimum, avoiding the need for highly trained experts to be on standby.
- **Safety and non-destructiveness:** The technique should pose minimal risk to patients and medical staff, and should preserve the surgical specimen to allow for later histopathological margin status confirmation.
- **Cost-effectiveness:** Implementation costs, including equipment and personnel, should be balanced against potential savings from reduced re-excision rates and fewer delays in adjuvant therapy.

2.4.1. Standard-of-care (SOC) techniques

The most commonly used SOC techniques for IMA include frozen sections, imprint cytology, intraoperative US, and specimen radiography [8, 28, 29, 71]. All techniques are further explained below. The diagnostic accuracy of these techniques, according to meta-analyses, is summarized in Table 2.1. For each technique, the results from the meta-analysis that included the largest number of studies are presented.

Frozen sections refer to a histopathology technique where a small piece of tissue is sampled from a resection specimen, is snap frozen, thinly sliced, mounted on a glass slide, and H&E stained to allow for microscopic analysis. While these slides are often of inferior quality compared to traditionally prepared slides, they allow for quick IMA. While frozen sections have a high diagnostic accuracy [71], the procedure is prone to sampling bias, is time-consuming, labour-intensive, and expensive as a pathologist has to be on standby to interpret the slides [8, 71]. Additionally, this process can damage specimen margins, interfering with final histopathology [71].

Imprint cytology is a histopathology technique in which freshly excised tissue is pressed onto a glass slide to transfer cells that can be studied under a microscope. While imprint cytology has a high diagnostic accuracy [8, 72], it suffers from similar disadvantages as frozen sections. The technique is prone to sampling bias, time-consuming, and logistically cumbersome [8]. In addition, only cells at the surface of the specimen can be examined, preventing determination of the margin width [8].

Intraoperative ultrasound (US) is an imaging technique that entails using an US probe to examine the resected tissue specimen and the surgical cavity. Compared with frozen sections or imprint cytology, intraoperative US shows lower sensitivity for positive margin detection [8, 72]. Reported diagnostic accuracy varies widely, as performance is highly dependent on the operator's skills and experience. An in-room breast radiologist is required to accurately perform US imaging, which makes the technique expensive and logistically demanding [73].

Specimen radiography is an imaging technique that delivers a planar 2D X-ray projection image of an excised specimen. While in most hospitals the specimen must be transported to the radiology department for specimen radiography, mobile scanners are available that can be used directly inside the operating theatre, thereby causing minimal disruption to the clinical workflow [71]. However, the sensitivity of this technique for positive margin detection is found to be rather low [8, 72, 73].

2.4.2. Emerging imaging-based techniques

As discussed in the previous section, all current SOC IMA methods have several drawbacks. They are either too time-consuming, lack sufficient diagnostic accuracy, are too expensive, or too technically demanding and disruptive of the surgical workflow [28]. Therefore, we review the most widely studied emerging imaging-based techniques for IMA below. Because these techniques rely on different types of electromagnetic radiation, Figure 2.5 visualises the electromagnetic spectrum for clarification. Table 2.1 shows the diagnostic performance of techniques that were studied in meta-analyses. Although many other technologies are under development, clinical evidence supporting their use for IMA during BCS is scarce. Techniques not further discussed include specimen MRI, Cerenkov luminescence imaging, flexible autoradiography, fluorescence imaging, mass spectrometry, and gamma- and beta-probes.

Tomosynthesis and Micro-CT reconstruct 3D images from 2D X-ray projection images (i.e., radiographies) taken at different rotation angles around the specimen. In case of tomosynthesis, an emulated 3D image is reconstructed from several 2D projection images taken across a limited rotation angle, whereas micro-CT images are reconstructed from 2D projection images systematically sampled across a 180° rotation angle, resulting in improved image quality. The principles behind CT imaging are explained in more detail in Section 2.5.1.1. While compact and mobile devices exist that allow for fast tomosynthesis or micro-CT imaging inside the operating theatre with minimal disruption of the surgical workflow, images have a poor soft tissue contrast. The sensitivity of these techniques is found to be higher compared to specimen radiography, but the performance is still lower than histopathological techniques [72, 73].

Optical spectrometry and imaging techniques distinguish tissue types based on how they interact with optical (visible, ultraviolet, and infrared) light. **Diffuse reflectance spectrometry** analyses elastically scattered (Rayleigh) light, in which the wavelength is preserved while the intensity changed due to tissue absorption. Light intensity variations reflect differences in tissue composition and morphology. When cameras are used to simultaneously capture spectral information across wider tissue regions, the technique is also referred to as hyperspectral imaging. **Raman spectrometry** analyses inelastically scattered light, in which wavelength shifts arose due to energy exchange with molecular vibrations in the tissue. Despite requiring longer acquisition times due to their weaker nature compared to Rayleigh signals, Raman signals provide more detailed information on chemical composition and molecular bonds. **Optical coherence tomography** generates cross-sectional images by measuring interference between reflected light and a reference beam, offering limited depth-resolved visualisation of microscopic tissue structures. All three techniques have been implemented in compact and mobile devices and have demonstrated good diagnostic performance [74–76], generally outperforming intraoperative US and specimen radiography. Whereas diffuse reflectance spectrometry and optical coherence tomography offer rapid imaging, Raman spectrometry is comparatively slow. Important limitations for all three methods include a restricted field-of-view, sampling bias, and the need for experts to interpret the complex spectral images. Moreover, diffuse reflectance and Raman spectrometry do not provide margin width information, while optical coherence tomography offers only limited depth information. Another related optical technique is confocal microscopy, which allows high-resolution microscopic surface imaging. However, the technique shares similar shortcomings and has limited clinical evidence [77].

Radiofrequency spectrometry involves exposing tissue to radiofrequency signals and distinguishing between benign and malignant tissue by analysing the reflected spectra, which vary according to the electrical impedance of the tissue. Hand-held probes were developed to enable real-time surface imaging using radiofrequency spectrometry. Important disadvantages of the technique include its labour-intensive nature, susceptibility to sampling bias, and inability to provide information on margin width. In addition, its performance is strongly operator dependent, hence the strong variations in reported diagnostic performance [72]. Bio-impedance spectrometry is a related technique that quantifies tissue impedance by transmitting electrical currents through the tissue. Clinical evidence for this technique is scarce [78].

Table 2.1. Diagnostic accuracy of standard-of-care intraoperative margin assessment techniques and emerging technologies according to meta-analyses [71–76].

Technique	N studies	Sensitivity [95% CI]	Specificity [95% CI]
Frozen sections [71]	19	0.81 [0.79 - 0.83]	0.97 [0.97 - 0.98]
Imprint cytology [72]	13	0.92 [0.77 - 0.98]	0.95 [0.90 - 0.97]
Intraoperative ultrasound [72]	7	0.72 [0.47 - 0.88]	0.87 [0.73 - 0.95]
Specimen radiography [72]	20	0.39 [0.24 - 0.56]	0.84 [0.77 - 0.89]
Tomosynthesis [73]	4	0.67 [0.53 - 0.78]	0.76 [0.66 - 0.84]
Micro-CT [73]	4	0.68 [0.34 - 0.90]	0.69 [0.54 - 0.81]
Diffuse reflectance spectrometry [74]	13	0.84 [0.78 - 0.89]	0.87 [0.79 - 0.91]
Raman spectrometry [75]	16	0.97 [0.92 - 0.99]	0.96 [0.91 - 0.98]
Optical coherence tomography [76]	18	0.91 [0.88 - 0.93]	0.88 [0.83 - 0.92]
Radiofrequency spectrometry [72]	3	0.73 [0.26 - 0.95]	0.53 [0.30 - 0.75]

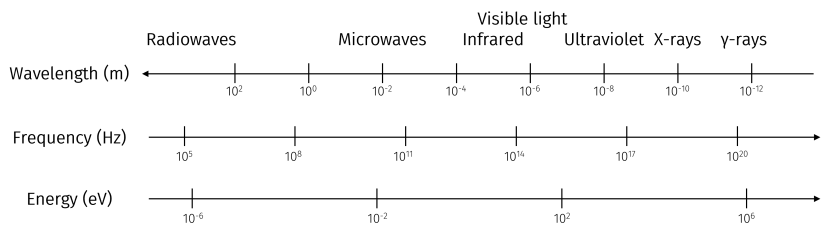


Figure 2.5. Electromagnetic spectrum. Figure based on [79].

2.5. Micro-PET-CT specimen imaging as IMA technique

As no ideal solution for IMA is currently available, XEOS has developed a compact and mobile high-resolution (micro-)PET-CT specimen scanner, called the AURA® 10. This system can be used inside the operating theatre to provide 3D micro-PET-CT images of excised lumpectomy specimens within approximately 10 min. The structural CT image allows precise delineation of the anatomical specimen margins, while the functional PET image allows detection of metabolically active tumour tissue. Together, these modalities provide comprehensive 3D information on margin status and margin width without sampling bias. Although the technique requires preoperative radiotracer administration, it has been shown to be safe and feasible in routine clinical practice [31]. Furthermore, micro-PET-CT specimen imaging has shown reliable intraoperative visualisation of different BCa subtypes, and has outperformed SOC IMA methods, offering clinically meaningful improvements in IMA during BCS [32, 33]. The following sections first introduce the fundamental principles of PET and CT imaging. Afterwards, a detailed overview of the clinical evidence supporting the feasibility, safety, and diagnostic accuracy of micro-PET-CT specimen imaging as IMA technique in BCa is provided.

2.5.1. Fundamental principles of PET-CT

Computed tomography (CT) provides 3D structural images of tissue anatomy, whereas positron emission tomography (PET) produces 3D functional images that visualise molecular tissue processes. The combined use of PET-CT has become widely adopted in oncology [80, 81]. PET enables early detection of cancer as it is highly sensitive to molecular changes, which precede structural changes [82]. However, the spatial resolution of CT images is superior, allowing precise anatomical localisation [82]. Therefore, the combination of these complementary modalities is essential, and combined PET-CT scanners are often built for easy and optimal image co-registration or alignment. In this section, we discuss the basic physical principles underlying CT and PET imaging, essential components of CT and PET scanners, and image reconstruction methods for both modalities. We limit our discussion to the core principles of both techniques. For more detailed information we refer the reader to the extensive work by Buzug [83], Phelps [84], and Lin [82].

2.5.1.1. Computed tomography (CT)

CT provides structural cross-sectional images of an object, based on the attenuation of X-rays passing through the object. X-rays are generated in an X-ray tube, attenuated as they travel through the object according to tissue-specific linear attenuation coefficients, and measured after transmission by scintillation detectors. Acquired projection data are organised into sinograms, from which the 3D distribution of tissue radiodensities is reconstructed, commonly expressed in Hounsfield units (HUs).

X-ray tubes are used to produce X-rays for CT imaging. Figure 2.6 provides an illustration of an X-ray tube. An X-ray tube is a vacuum tube with a cathode and (rotating) anode at both ends [83]. The cathode contains a heated filament from which electrons are emitted. These electrons are accelerated towards the anode by the accelerating voltage applied across cathode and anode. When the electrons enter into the anode material, they are decelerated and X-ray radiation is produced. A spectrum of X-rays with a range of energies will be emitted. The energy range is defined by the applied acceleration voltage, while the tube current defines the intensity of the X-ray beam. The X-ray energy spectrum consists of a continuous distribution caused by bremsstrahlung, superimposed by a characteristic line spectrum that depends on the anode material, e.g. tungsten [83]. The anode is often rotated to prevent it from overheating as it is hit by the electron beam.

X-ray attenuation takes place as the X-rays travel through the object that is being imaged. As X-rays travel through material, their flux

will exponentially decay through absorption and scattering [84]. This exponential decay of the flux (I) is described by Equation 2.1, where x indicates the distance travelled through the material, and μ represent the **linear attenuation coefficient** of the material, indicating the probability per unit distance that an interaction will occur [84]. Different tissues have different linear attenuation coefficients. Higher atomic numbers and higher physical density of tissues, result in higher attenuation coefficients [84]. The goal of CT imaging is to reconstruct the distribution of tissue attenuation coefficients.

$$I(x) = I(0) e^{-\mu x} \quad (2.1)$$

Scintillation detectors are used to measure the flux of the attenuated X-rays escaping from the imaged object. Nowadays, the most commonly used X-ray detectors in CT scanners are solid-state scintillation detectors, as illustrated in Figure 2.6 [83]. Such a detector consists of two main parts: a scintillation crystal that converts incoming X-rays into visible light photons and a photodiode that converts the visible light into an electrical current proportional to the intensity of the X-ray radiation [83]. The detector typically integrates the total electrical charge generated by incident X-rays over a predefined time window, after which the signal is read out and converted into a digital output. An appropriate scintillation material like caesium iodide (CsI) has a high efficiency and speed for the conversion of X-rays into visible light [83]. To improve image quality, an antiscatter collimator or grid can be placed on top of the scintillation crystal [83]. Such collimator filters out scattered photons that are not travelling along the line of sight between X-ray source and detector, thereby improving image contrast. However, an important drawback of a collimator is that it absorbs part of the primary radiation, necessitating a higher radiation dose to maintain adequate signal levels.

Sinograms are matrices collecting all acquired CT imaging data [82, 83]. The X-ray source and detector are placed on opposite sides of the object of interest, and rotate around it to make projection images at different angles. By comparing the ingoing and outgoing flux of X-rays at every angle, the integral of the linear attenuation coefficients along the current projection line can be estimated. To be able to reconstruct tomographic or cross-section images, X-ray projection images have to be sampled across an angle of at least 180°. All measured integrals of attenuation coefficients can be organised in a matrix according to the current rotation angle and distance from the rotation centre. Such matrix is called a sinogram as one point in the imaging field will describe a sinusoid curve in this matrix.

Image reconstruction methods are required to reconstruct an estimate of the 3D distribution of linear attenuation coefficients in the imaged tissue from the X-ray projection sinograms. The most common classical reconstruction method for CT is **filtered backprojection (FBP)** [83]. FBP is a fast analytical approach that provides an exact image reconstruction of tomographic data in case of noiseless, fully sampled data. In CT, high photon statistics and dense sampling make analytical reconstruction feasible and computationally efficient. While computationally more expensive, iterative reconstruction methods can often improve image quality. Common examples include **algebraic reconstruction techniques (ARTs)**, with variants such as the simultaneous iterative reconstruction technique (SIRT) and the simultaneous algebraic reconstruction technique (SART), as well as multiplicative update methods such as the image space reconstruction algorithm (ISRA), and more advanced model-based iterative reconstruction techniques [82]. Iterative methods allow the mathematical model underlying the reconstruction to better match the real physical acquisition process, facilitating incorporation of physical effects and correction mechanisms to reduce image artefacts. Iterative methods are especially beneficial in low-dose imaging, but are increasingly applied across a broad range of dose levels in modern CT scanners, while FBP remains attractive for its speed and simplicity [83].

Artefact corrections can be performed to improve the accuracy of the reconstructed images. Common artefacts addressed during reconstruction include metal artefacts and beam-hardening artefacts [83]. Metal artefacts typically arise from metallic objects with extremely high attenuation, which can cause photon starvation and consequently result in streaking and signal voids in the reconstructed image. Beam-hardening artefacts result from the polychromatic nature of the X-ray spectrum. Lower-energy photons are more strongly attenuated, increasing the mean energy of the transmitted beam and introducing non-linearities and image distortions. These beam-hardening artefacts are particularly pronounced in the presence of highly attenuating structures, such as bone.

Hounsfield units (HUs) are dimensionless quantitative units commonly used to express the reconstructed radiodensity of tissues in CT images [83]. As expressed in Equation 2.2, HUs express the linear attenuation coefficient of a tissue relative to the attenuation coefficient of water. Radiodense structures like bone will appear white in the CT images, while areas with low radiodensity like air will appear black [85]. Evidently, water has HUs of 0, while bone has HUs of +1000 and higher, and air has HUs around -1000 [85].

$$\mu[\text{HU}] = 1000 \cdot \frac{\mu - \mu_{\text{water}}}{\mu_{\text{water}}} \quad (2.2)$$

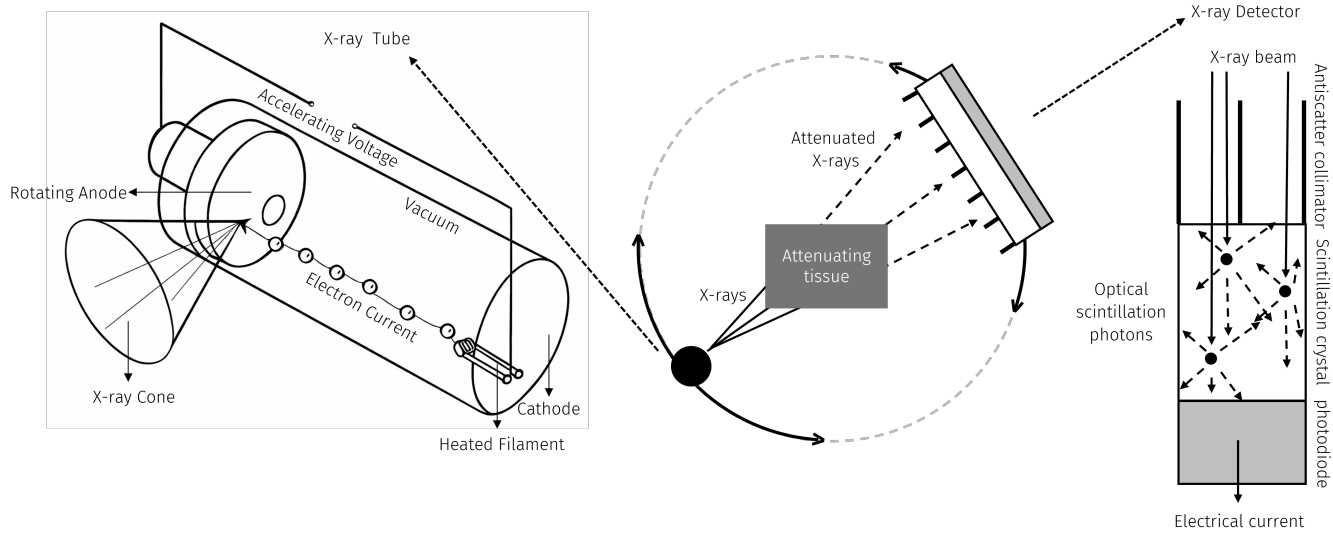


Figure 2.6. Basic principles of CT imaging. In CT, X-rays produced in an X-ray tube are sent through an object, and the attenuated signal is recorded by detectors opposite of the source. As the X-ray tube and detector rotate around the object, multiple projection images are acquired that are later reconstructed into tomographic images representing the spatial distribution of tissue attenuation coefficients. Figures based on [83].

2.5.1.2. Positron emission tomography (PET)

For PET imaging, positron-emitting radiotracers are administered to the patient to visualise biological processes. Radioactive decay of these tracers leads to positron-electron annihilation, producing pairs of 511 keV gamma photons. Scintillation detectors are used to measure coincidences of these gamma photons. Acquired events are organised into sinograms and reconstructed to obtain the 3D activity distribution in the tissue, commonly expressed using semi-quantitative standardised uptake values (SUVs).

PET radiotracers are intravenously administered to the patient to allow visualisation of specific biological processes in the body. A radiotracer consists of two main components: a ligand and a radionuclide [82]. The ligand is a molecule of biological interest of which we want to image the distribution in the body. The radionuclide is a radioactive isotope coupled to the ligand allowing to visualise the ligand of interest. In case of PET imaging, positron-emitting radionuclides are used [82]. These radionuclides, produced in a cyclotron or generator, have an unstable highly-energetic nucleus in which a proton (p^+) can decay into a neutron (n), thereby emitting a positron (e^+) and an electron neutrino (ν_e), as indicated in Equation 2.3. This decay reaction is known as β^+ decay. Emitted positrons will quickly combine with a nearby electron. During this reaction, the mass of both particles is converted into energy in the form of two 511 keV gamma rays that travel in opposite directions [82]. This **positron-electron annihilation** is visualised in Figure 2.7. Strictly speaking, the emission angle of the gamma photons follows a Gaussian distribution around 180° with full width at half maximum (FWHM) of 0.5° , an effect known as photon acollinearity that limits the spatial resolution of PET [84]. Gamma photons undergo less attenuation than X-rays and are therefore more likely to escape from the body where they can be detected, enabling reconstruction of the radiotracer distribution in PET.



Common PET radionuclides include $^{18}_9\text{F}$ (decaying to $^{18}_8\text{O}$) and $^{68}_{31}\text{Ga}$ (decaying to $^{68}_{30}\text{Zn}$) [84]. Two important characteristics of PET radionuclides are the positron range and the half-life [84]. The positron range defines the distance across which an emitted positron can travel before annihilation takes place. It depends on the positron energy and the tissue type. This range limits the achievable spatial resolution for a PET image. For $^{18}_9\text{F}$, the positron range is relatively short (typically ≤ 1 mm in most human tissues), whereas it is longer for $^{68}_{31}\text{Ga}$ [82, 84]. The half-life (τ) represents the time required for half of all nuclei to decay, considering the decay of radionuclides is a stochastic process. The half-life should be long enough to allow for imaging, while short enough such that the patient does not stay radioactive for too long. The half-life of $^{18}_9\text{F}$ is 109.8 min, while the half-life

of $^{68}_{31}\text{Ga}$ is 67.6 min [84]. Radiotracer activity is often expressed in becquerel (Bq), where 1 Bq indicates 1 decay per second [82]. The radiotracer activity (A) decays exponentially over time (t), as described in Equation 2.4.

$$A(t) = A(0) e^{-t \frac{\ln 2}{T}} \quad (2.4)$$

Common PET radiotracers include both widely used metabolic tracers and more targeted tracers, tailored to specific applications. By far the most widely available and most commonly used radiotracer in oncological PET imaging is [^{18}F]FDG [82]. It is a radioactive sugar-analogue that gets accumulated in metabolically active cells, such as tumour cells [83, 86]. An important disadvantage of [^{18}F]FDG is its tendency to show non-specific uptake in non-malignant metabolically active regions, such as inflamed tissue [87]. Beyond [^{18}F]FDG, several more targeted radiotracers are being developed. We here focus on targeted tracers for BCa. As the majority of BCas are ER-positive, the ER-binding radiotracer [^{18}F]fluoroestradiol ([^{18}F]FES) shows promising first results [56]. In addition, HER2-targeting radiotracers based on the monoclonal antibodies trastuzumab and pertuzumab, including [^{89}Zr]pertuzumab and [^{89}Zr]trastuzumab, are currently under investigation [88].

Scintillation detectors are used to detect coincident gamma photons produced during electron-positron annihilation. As illustrated in Figure 2.7, a PET scanner consists of a full ring of detector blocks. Each detector block typically contains a scintillation crystal that converts incident 511 keV gamma photons into visible or near-ultraviolet light, coupled to a photodetector that transforms this light into an electrical signal [89]. When a gamma photon interacts in the scintillation crystal, its energy is converted into a short burst of visible light photons [82]. The most commonly used scintillation material in modern PET scanners is lutetium yttrium oxyorthosilicate (LYSO), due to its high stopping power, short decay time, favourable energy resolution, and high light output [82]. The emitted optical light is then detected by photomultipliers, which convert it into an amplified electrical signal proportional to the total deposited photon energy. Two main types of photomultipliers are used: photomultiplier tubes (PMTs) and silicon photomultipliers (SiPMs) [82]. PMTs are vacuum tubes which convert the incoming visible light into electrons that are accelerated and amplified to produce an electrical current. SiPMs are silicon-based solid-state detectors, offering advantages in size, cost, and performance compared to PMTs. Coupling multiple photomultipliers to a scintillation crystal allows measuring the spatial spread of scintillation photons and enables more precise localisation of the coincidence location of a gamma ray.

Sinograms collect all coincidences detected during PET acquisition [84]. A coincidence detection occurs when two photons with 511 keV energy are detected at the same time. It is assumed that the original annihilation event took place along the line connecting the two detected photons, called the **line of response (LOR)** [82]. Accurate localisation of the annihilation point is limited by detector physics, the positron range, and photon acollinearity, which together limit the scanner's spatial resolution. To organise all detected LORs, they are collected in sinograms, where each LOR is recorded based on the distance of the LOR from the centre of the detector ring and the angle of the LOR.

Image reconstruction methods are required to reconstruct an estimate of the 3D radiotracer distribution in the imaged tissue from the acquired sinograms. The most common reconstruction methods are iterative statistical reconstruction methods like **maximum likelihood expectation maximisation (MLEM)**, and **ordered subset expectation maximisation (OSEM)** [82]. Since PET consists of low-count measurements, analytical approaches such as FBP tend to amplify noise and may result in suboptimal image quality [82]. Iterative methods such as MLEM and OSEM are therefore more suited for PET, as they explicitly model the statistical nature of the data and allow incorporation of physical effects and correction factors into the reconstruction process to reduce artefacts [82]. OSEM generally converges faster than MLEM by using subsets of the acquired sinogram data, and OSEM with a single subset is mathematically equivalent to MLEM.

Artefact corrections can be performed to improve accuracy of the reconstructed activity images. One of the most important corrections is attenuation correction [82]. As gamma photons travel through tissue, they are attenuated depending on the path length and tissue density. Therefore, the sinograms should be corrected for attenuation. The corresponding CT image is typically used to estimate the attenuation along each LOR. Another common correction is scatter correction. Gamma photons may change direction when they undergo scattering inside the patient, causing events to be assigned to incorrect LORs. This introduces background signal and degrades image contrast. Most scatter correction techniques use the attenuation map to estimate the scatter distribution, which can be subtracted from the measured sinograms. Furthermore, post-processing techniques may be applied to reduce image noise [82]. For example, median filtering can be used to suppress noise while preserving structural details.

Standardised uptake values (SUVs) are dimensionless semi-quantitative units commonly used to express radiotracer uptake in PET images [84]. While reconstructed PET images are typically reported in Bq/ml, comparison across patients or scans requires normalisation of the measured activity concentration. SUVs normalise the measured activity to the administered dose and to the patient's body weight and account for the

time between injection and imaging. SUVs can be interpreted as the ratio of tissue activity concentration to the average whole-body activity concentration. The standard formula, given in Equation 2.5, assumes a uniform distribution of injected activity and a tissue density of one kg/l, making SUV a dimensionless quantity. Measured activity values must be decay-corrected to the time of injection. A common strategy to delineate lesions in PET images is to use an isocontour at a fixed percentage (e.g., 40%) of the maximum uptake within the lesion (SUV_{max}) [82, 90]. Another widely used approach sets the threshold adaptively according to the tumour-to-background ratio, which compensates for varying levels of surrounding background activity that a fixed percentage of SUV_{max} ignores [91].

$$SUV = \frac{\text{measured activity } [\frac{\text{kBq}}{\text{ml}}]}{\frac{\text{injected activity}[\text{MBq}]}{\text{patient's body weight } [\text{kg}]}} \quad (2.5)$$

Lean body mass SUVs are used to reduce the influence of body composition on SUV values. Most radiotracers, such as [^{18}F]FDG, are hydrophilic and do not accumulate in adipose tissue. Consequently, normalisation by total body weight may overestimate SUVs in heavier patients with higher body fat percentages. To overcome this bias, lean body mass (LBM) can be substituted for total body weight in Equation 2.5. LBM expresses the patient's mass of non-adipose tissue. A common formulation to estimate LBM from the patient's body weight (BW in kg) and height (H in cm), is the **Janmahasatian formulation** [92]. Equations 2.6 and 2.7 show the Janmahasatian formulation of LBM for men and women, respectively. Separate formulas are required due to sex-related differences in body fat percentages. Although more advanced normalisation approaches exist, they are rarely used in routine practice. SUVs remain semi-quantitative, as multiple biological and technical factors affect their values, often resulting in poor reproducibility between centres.

$$LBM_{\text{men}} = \frac{9.27 \cdot BW}{6.68 + 2160 \cdot (\frac{BW}{H^2})} \quad (2.6)$$

$$LBM_{\text{women}} = \frac{9.27 \cdot BW}{8.78 + 2440 \cdot (\frac{BW}{H^2})} \quad (2.7)$$

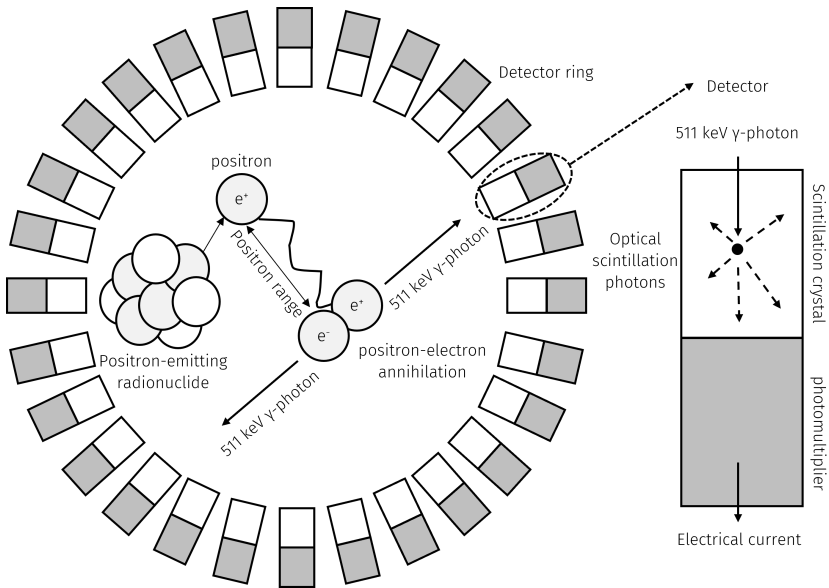


Figure 2.7. Basic principles of PET imaging. In PET, an injected positron-emitting radiotracer leads to positron-electron annihilation reactions that produce two gamma photons travelling in opposite directions. These gamma photons are picked up by a ring of detectors, allowing later reconstruction of the radiotracer distribution in the imaged tissue. Figures based on [84].

2.5.2. Rationale and system design for *ex vivo* micro-PET-CT

While *in vivo* PET-CT imaging has become widely adopted in clinical practice, XEOS is the first to build a high-resolution (micro-)PET-CT scanner dedicated to *ex vivo* specimen imaging. The structural CT image can be used to delineate the anatomical margins of the specimen, while the functional PET image is used to detect metabolically active tumour tissue. Figure 2.9 shows examples of micro-PET-CT images of two BCa specimens. The greyscale image expressed in HUs represents the micro-CT image. Darker adipose tissue can be distinguished from lighter connective, healthy glandular, and (pre)malignant tissue. The overlaying colour image expressed in SUVs represents the $[^{18}\text{F}]\text{FDG}$ micro-PET image. The micro-PET image allows to detect whether highly metabolic tissue is present at the border of the specimen, indicating a high risk on a positive margin. This IMA technique can visualise the entire specimen in 3D, allowing to assess all margins and the margin width without sampling bias.

While the fundamental principles underlying *in vivo* and *ex vivo* PET-CT imaging are the same, scanner designs differ. For specimen imaging, a compact and mobile scanner was developed that can be positioned

inside the operating theatre, making images immediately available for the surgeon. Furthermore, the specimen scanner has a smaller field-of-view and is fully enclosed within a shielded housing to ensure minimal radiation exposure to device operators. In addition, during CT imaging, the specimen is rotated while the X-ray source and detector remain stationary, improving mechanical stability of the system.

Compared with *in vivo* imaging, *ex vivo* imaging offers several advantages that enable higher spatial resolution. The absence of respiratory motion eliminates motion-related blurring. For CT imaging, higher X-ray doses can be used without concern for patient radiation risk. For PET imaging, detectors can be positioned in close proximity to the specimen, resulting in a high detection sensitivity. The smaller scanner diameter also reduces the impact of photon acollinearity. Spatial resolution is conventionally expressed as the FWHM of the Gaussian-like point spread function representing an ideal point source in the reconstructed image. The micro-PET module of the AURA[®]10 scanner has demonstrated the ability to resolve submillimetre features. Figure 2.8 shows a micro-PET image of a Derenzo phantom filled with [¹⁸F]FDG, in which the 1 mm rods can clearly be distinguished, and even the 0.9 mm and 0.8 mm rods remain discernible, corresponding to FWHM values down to 0.8 mm. It should be noted, however, that post-processing techniques such as median filtering can affect the measured spatial resolution. Moraitis et al. [35] reported values of 1.77-1.67 mm for [¹⁸F] in median-filtered AURA[®]10 micro-PET images. While it is common for *in vivo* scanners to have a higher in-plane resolution compared to axial resolution, the resolution of the AURA[®]10 is more isotropic. Improved spatial resolution mitigates partial volume effects, leading to sharper images and more accurate SUV quantification in small lesions.

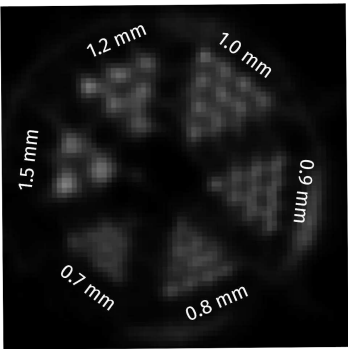


Figure 2.8. Micro-PET image of a Derenzo phantom filled with [¹⁸F]fluorodeoxyglucose ([¹⁸F]FDG) acquired with the AURA[®]10 from XEOS.

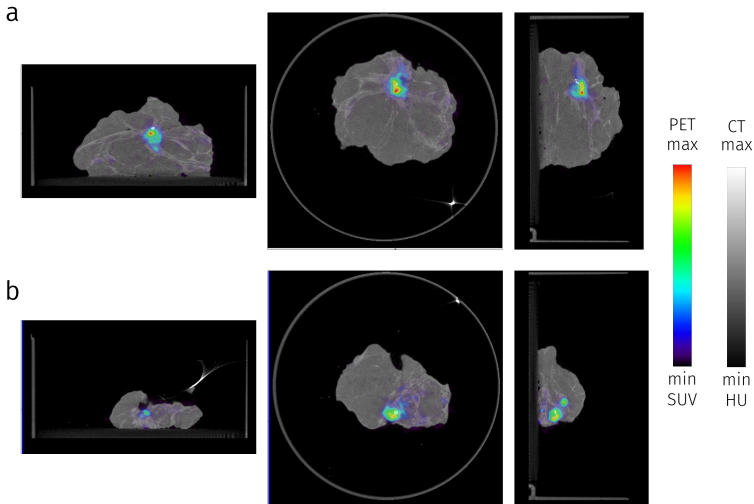


Figure 2.9. Two examples of micro-PET-CT images of breast cancer specimens. Three orthogonal views are shown for each specimen. (a) Specimen with histopathology-confirmed negative resection margin. (b) Specimen with histopathology-confirmed positive resection margin.

2.5.3. Feasibility and safety of intraoperative implementation

The workflow for IMA during BCS using micro-PET-CT specimen imaging is illustrated in Figure 2.10. Before surgery, the patient receives an intravenous injection with a radiotracer, which is often [^{18}F]FDG in case of BCa. Then, the patient undergoes BCS, and immediately after excision the BCa specimen is placed inside the micro-PET-CT scanner for visualisation. Within approximately 10 min (comparable to the duration of the sentinel node procedure), a 3D fused micro-PET-CT image of the lumpectomy specimen is available. This reconstructed image can be interpreted by the surgeon or another physician to determine the margin status. As the technique is non-destructive, the specimen can still undergo subsequent routine histopathological analysis. For optimal micro-PET-CT image quality, acquisition is ideally performed 1-3 h after [^{18}F]FDG injection. [^{18}F]FDG uptake in most lesions increases during the first two hours post-injection before reaching a plateau. Therefore, cutting off the blood supply to the tumour too early may reduce image contrast [82]. On the other hand, delayed imaging may also compromise image quality due to radiotracer decay.

Lambert et al. [31] published a feasibility study in 2025 evaluating the safety and feasibility of implementing this workflow for intraoperative micro-PET-CT specimen imaging. The technique was applied in 32 surgeries across various applications (e.g., (para)thyroid surgery, urological surgery,

oncological procedures such as BCa surgery, etc.) using different types of radiotracers. Radiation dose rates were monitored throughout the procedures. The median absorbed dose per procedure, normalized to an injected activity of 1 MBq/kg, was 15.6 μ Sv for surgeons and 14.1 μ Sv for instrumenting nurses. These values are well below the dose limit for occupationally exposed staff of 20 mSv per year. The study further demonstrated that high-quality images could be obtained at a dose level of 1 MBq/kg, owing to the high sensitivity of the PET detectors, ensuring safe and feasible clinical implementation.

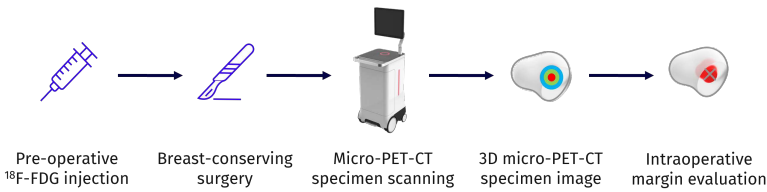


Figure 2.10. Workflow for intraoperative micro-PET-CT imaging of lumpectomy specimens. The figure contains an illustration of the AURA[®] 10 micro-PET-CT specimen scanner by XEOS.

2.5.4. Diagnostic performance for IMA in breast cancer

In this section, the current clinical evidence on the diagnostic accuracy of micro-PET-CT for IMA in BCS is summarized. Two clinical trials were dedicated to this performance evaluation: the BIMAP and BrIMA trials. Results of both studies are described in the following sections.

2.5.4.1. BIMAP

A first proof-of-concept study, titled “Intra-operative PET-CT: a Novel Approach to Determine Excision Margins in Lumpectomy Breast Cancer”, later called the BIMAP (Breast Intraoperative Margin Assessment Pilot) study, was initiated in 2018 (ClinicalTrials.gov identifier NCT04343079). This clinical open-label, prospective, single-arm, interventional study enrolled patients with early-stage BCa scheduled for BCS at Ghent University Hospital and had the goal of exploring the feasibility of micro-PET-CT specimen imaging as a new IMA technique in BCa. The diagnostic performance and clinical value of the technique were assessed by comparing with final histopathology as gold standard, as well as with the SOC IMA technique of specimen radiography applied at the Ghent University Hospital. As this was a feasibility study, no intraoperative decisions were taken based on the acquired micro-PET-CT images.

Göker et al. [30] published intermediate results for the first 20 patients enrolled in this study in 2020. Patients included 15 NST, 4 ILC, and 1 DCIS case, and patients were preoperatively injected with 4 MBq/kg [^{18}F]FDG. All micro-PET images were reconstructed to emulate a 10 min acquisition with an injected dose of 0.8 MBq/kg. A region of high [^{18}F]FDG uptake was visualised in each lumpectomy specimen, which corresponded to the tumour as described by histopathological evaluation. All images were retrospectively analysed by three breast surgeons, who predicted the margin status for all cases. Comparing the predicted with the true histopathological margin status for all 20 cases, resulted in a sensitivity and specificity of 79% and 72%, respectively, averaged across the three observers. Image assessment took around 1 min per patient, and considerable inter-reader variability was observed. Image interpretation was found to be the most challenging for ILC cases, possibly due to their infiltrative growth pattern resulting in lower tumour volume and lower [^{18}F]FDG uptake. Overall, this intermediate analysis showed promising results for micro-PET-CT imaging to facilitate IMA during BCS.

More extensive results were later published in a follow-up study by De Crem et al. [32] in 2025. In total, 49 patients were enrolled in the study, of which 41 were found eligible for further analysis. Patients included 31 NST, 6 ILC, 2 DCIS, and 2 neoadjuvant-treated (NAT) cases. The 2 NAT cases in this study were NST cases pretreated with NAC. Table 2.2 provides an overview of all studied lumpectomy specimens, together with their true margin status according to histopathology. For 25 patients, specimen radiography was performed as SOC IMA technique. While the first study patients received an injection with 4 MBq/kg [^{18}F]FDG, this dose was later adjusted to 0.8 MBq/kg [^{18}F]FDG. For uniformity, all micro-PET images were reconstructed to emulate a 10 min acquisition with an injected dose of 0.8 MBq/kg. All images were postoperatively assessed by seven physicians with varying levels of experience (three breast surgeons, one gynaecology resident, two nuclear medicine physicians, and one radiologist). All observers scored the margins as positive, close (≤ 1 mm), or negative, and results were compared to histopathology and SOC. A more strict interpretation workflow was followed compared to the first intermediate analysis reported by Göker et al [30]. In this study the same “ink on tumour” definition of a positive margin was used for all BCa types, and in case of invasive tumours, only the invasive tumour component was accounted for. Best results were obtained when considering close margins as positive. The average sensitivity and specificity for positive margin detection across the seven physicians when considering close margins as positive, are shown in Table 2.2 per BCa type.

For the combined group of all BCa subtypes, the seven physicians reached an average sensitivity and equal specificity of 81%. For the main focus group of NST cases, the seven physicians reached an even higher average sensitivity and specificity of 91% and 86%, respectively. For the NST

tumours, 9/31 cases were found to have a positive margin according to histopathology. Of these positive cases, 4/9 were identified through SOC, and 8/9 were identified by all seven physicians on micro-PET-CT specimen images. When comparing the performance to other IMA techniques in Table 2.1, it is observed that micro-PET-CT specimen imaging in a combined group of various BCa subtypes is outperforming specimen radiography, tomosynthesis imaging, micro-CT imaging, and radiofrequency spectrometry. When looking at the main focus group of NST tumours, it can be seen that only imprint cytology and Raman spectrometry are outperforming micro-PET-CT specimen imaging. In addition, interpretation results for the NST group revealed only minimal differences in performance of physicians with different levels of experience, suggesting even inexperienced physicians can achieve high diagnostic performance after interpretation training. While ILC, and NAT cases were more challenging to interpret, the dataset for these cases was small, and individual results should not be overanalysed. In conclusion, it was shown that micro-PET-CT specimen imaging can facilitate IMA during BCS, potentially reducing re-operation rates in the future.

2.5.4.2. BrIMA

In 2022, the BrIMA (Breast Intraoperative Margin Assessment) trial was initiated (Clinicaltrials.gov identifier NCT04999917). The BrIMA trial is a multicentre, open-label, prospective, single-arm, interventional study. This study, titled “Specimen PET-CT imaging for Intraoperative Margin Assessment in Breast Cancer”, enrolled early-stage BCa patients scheduled for BCS across six European BCa centres in Belgium (n=3), Germany (n=2), and Italy (n=1). Study results were reported by Göker et al. [33]. The primary objective of this study was to evaluate the success rate of micro-PET-CT specimen imaging in addressing positive margins for the invasive component in NST BCa specimens. In addition, other BCa subtypes were also studied as secondary objectives. The most important difference with the BIMAP study, is that the BrIMA study collected a more extensive dataset from a larger, more diverse, and more representative patient population across different countries following a more stringent study design. Whereas the BIMAP study was conducted in an exploratory research setting, the BrIMA study integrated micro-PET-CT imaging directly into the clinical workflow, enabling surgeons to immediately act upon the imaging results.

In total, 206 patients were included in the study, of which 152 were found eligible for further diagnostic performance analysis. Patients included 84 NST, 16 ILC, 13 DCIS, and 39 NAT cases. The 39 NAT cases in this study comprised 14 cases of NST pretreated with NAC, 16 cases of NST pretreated with aromatase inhibitors (NST-AI), and 9 cases of ILC pretreated with aromatase inhibitors (ILC-AI). An overview of all studied

lumpectomy specimens, together with their true margin status according to histopathology, can be found in Table 2.2. All patients received a preoperative injection with 0.8 MBq/kg [^{18}F]FDG and all specimen images were acquired with the AURA[®]10 (XEOS, Belgium), using a 10 min PET acquisition. An important difference with the proof-of-concept study was that all micro-PET-CT images were assessed intraoperatively by the surgeons, enabling them to decide on additional cavity shaves based on the imaging results. The “ink on tumour” definition of a positive margin was used for all BCa types, except for DCIS cases where the margin was considered positive when the minimal distance between the tumour and the resection border was ≤ 2 mm. We here discuss the results where only the invasive tumour component is taken into account in case of invasive BCa specimens, in correspondence with the primary study objective and with the BIMAP results. Image interpretations were compared to histopathology and SOC IMA techniques (i.e., specimen radiography, specimen US, and palpation). The sensitivity and specificity for positive margin detection reached by the surgeons intraoperatively interpreting the micro-PET-CT images, are shown in Table 2.2 per BCa type.

For the combined group of all BCa subtypes, the surgeons reached an average sensitivity and specificity of 77% and 67%, respectively. For the main study group of NST cases, the surgeons reached a sensitivity and specificity of 71% and 66%, respectively. For the NST tumours, 14/84 specimens were found to have a positive margin according to histopathology. Micro-PET-CT imaging identified 10/14 positive cases, while SOC only identified 3/14 cases. For all patients combined, McNemar’s test with Bonferroni correction showed that the success rate of intraoperatively addressing positive margins with micro-PET-CT specimen imaging was significantly higher compared to no use of IMA methods, and compared to SOC IMA techniques. For the group of NST cases, the success rate of intraoperatively addressing positive invasive margins also significantly improved when using micro-PET-CT specimen imaging compared to no use of IMA methods. However, in the NST group, the differences between micro-PET-CT specimen imaging and SOC IMA techniques were not found to be statistically significant.

While higher diagnostic performance was reported for several IMA technologies, as seen in Table 2.1, it is important to note that only few studies employ the same interventional, prospective, and multicentre study design as the BrIMA study. Notably, the diagnostic performance of micro-PET-CT specimen imaging was also higher in the BIMAP study [32]. Similar stringent study designs were applied by Schnabel et al. [93] who found a sensitivity of 75% and specificity of 46% for radiofrequency spectrometry, and by Zysk et al. [94] who found a sensitivity of 55%-65% and a specificity of 68%-70% for optical coherence tomography. The BrIMA study has shown that micro-PET-CT specimen imaging outperforms both methods. Overall, micro-PET-CT specimen imaging provides a good balance between

sensitivity and specificity, while offering important advantages like 3D imaging of the entire specimen and minimal disruption of the surgical workflow.

In addition, inter-reader reproducibility was assessed by asking four surgeons, trained for image interpretation, to independently re-interpret all images after enrollment was completed. The intraclass correlation coefficient (ICC) for the average of the four surgeons (ICC(2,4)) was 0.76, indicating good agreement for the averaged assessments and high reproducibility when interpretations were combined across observers. In contrast, the ICC for a single observer (ICC(2,1)) was 0.45, indicating poor to moderate agreement at the level of individual observers, reflecting variability in single-reader assessments. In conclusion, it was shown that micro-PET-CT specimen imaging provides reliable intraoperative visualisation across BCa subtypes and outperforms SOC IMA methods, offering clinically relevant improvements in IMA during BCS. The use of this integrated approach might lead to a substantial reduction of re-excision rates in BCS.

Table 2.2. Overview of specimens included in the BIMAP and BrIMA studies per breast cancer (BCa) subtype, as well as the performance of micro-PET-CT for margin assessment across subtypes. For each study (BIMAP and BrIMA) and BCa subtype, the number of analysed lumpectomy specimens (# specimens) and the number of specimens with histopathologically confirmed positive margins (# positives) are reported. For the BIMAP study, the mean sensitivity and specificity for positive margin detection, based on postoperative interpretation of micro-PET-CT images by seven physicians, are provided per BCa subtype [32]. For the BrIMA study, the sensitivity and specificity obtained from intraoperative image interpretation by physicians are reported [33].

Study:		BIMAP		
BCa type	# specimens	# positives	sensitivity	specificity
NST	31	9 (29%)	91%	86%
ILC	6	2 (33%)	50%	54%
DCIS	2	1 (50%)	100%	100%
NAT	2	1 (50%)	43%	43%
TOTAL	41	13 (32%)	81%	81%
Study:		BrIMA		
BCa type	# specimens	# positives	sensitivity	specificity
NST	84	14 (17%)	71%	66%
ILC	16	6 (38%)	83%	70%
DCIS	13	10 (77%)	80%	67%
NAT	39	5 (13%)	80%	68%
TOTAL	152	35 (23%)	77%	67%

NST = invasive carcinoma of no special type, ILC = invasive lobular carcinoma, DCIS = ductal carcinoma in situ, LCIS = lobular carcinoma in situ, NAT = neoadjuvant-treated carcinoma

3 | Fundamentals of deep learning and its application to medical image analysis

3.1. Introduction

In this dissertation, deep learning (DL) models are developed to support physicians in the interpretation of micro-PET-CT images of lumpectomy specimens. The first part of this chapter introduces the fundamental technical concepts underlying DL. The relationships between artificial intelligence (AI), machine learning (ML), and deep learning (DL) are clarified, followed by an overview of the main learning strategies: supervised, unsupervised, and reinforcement learning. Afterwards, deep artificial neural networks (ANNs) are introduced as the model architecture used in DL and their iterative training process is explained. This part is concluded by discussing regularisation techniques aimed at improving the generalisation performance of DL models.

In this work, DL models are employed to segment breast tumours in micro-PET-CT images, thereby allowing the classification of lumpectomy specimens according to their margin status. The second part of this chapter therefore focuses on the application of DL in computer vision tasks for medical image analysis. This part begins by distinguishing between classification and semantic segmentation tasks. Next, convolutional neural networks (CNNs) are introduced as a type of ANN that is particularly suited for image analysis, and the architecture of two widely used CNNs in computer vision are described: ResNet for classification and U-Net for segmentation. Afterwards, we discuss commonly applied metrics to evaluate model performance in both classification and segmentation problems. The final section of this part addresses specific challenges associated with DL in medical image analysis, along with emerging trends in the field.

Finally, we briefly introduce the key concepts of classical iterative image co-registration algorithms. In this dissertation, we obtain ground-truth breast tumour segmentations to create the DL models by co-registering micro-PET-CT images of lumpectomy specimens with the corresponding histopathology slides. Therefore, we provide the minimal background needed to understand the ground-truth label generation pipeline presented in the next chapter.

3.2. Fundamentals of deep learning

3.2.1. Artificial intelligence, machine learning, and deep learning

Artificial intelligence (AI) refers to the development of intelligent computer systems capable of mimicking human behaviour [95]. Over the last few decades, rapid advances in computational power, algorithms, and data availability have accelerated the development of this field [96]. Early AI systems were primarily knowledge-based systems, relying on explicitly programmed rules to perform specific tasks. **Machine learning (ML)**, a subfield of AI, enables systems to learn from example data or past experience how to perform a specific task rather than relying on predefined rules [96]. By identifying patterns in data, ML models can make predictions or decisions, often capturing complex relationships that may not be readily apparent to human observers. **Deep learning (DL)** is a subfield of ML that focuses on a specific type of models, namely deep artificial neural networks (ANNs) [97]. These models are particularly powerful for tasks that require detecting complex patterns, such as (medical) image analysis and natural language processing. In this dissertation, DL models are applied to analyse micro-PET-CT images of lumpectomy specimens. Despite their strong performance, DL models typically require large training datasets and are often considered “black boxes” as they lack interpretability [96]. The relationship between AI, ML, and DL is shown in Figure 3.1.

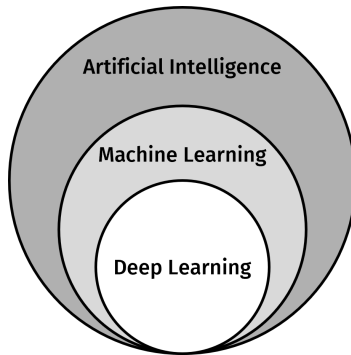


Figure 3.1. Relation between artificial intelligence, machine learning, and deep learning.

3.2.2. Supervised, unsupervised, and reinforcement learning

When training an ML or a DL model to perform a specific task, three main learning strategies can be applied [95, 96]:

- **Supervised learning** is a learning approach in which a model infers knowledge from labelled data consisting of known input-output pairs. The model learns the relationship between inputs and their corresponding outputs by minimising the difference between its predictions and the ground-truth labels. Once trained, the learned relationship is used to make predictions for new, unseen data.
- **Unsupervised learning** involves learning associations in data without the availability of labelled examples. The aim is to find hidden structures and relations in input data, by for example clustering the data into similar groups.
- **Reinforcement learning** is often used in gaming or robot control. An artificial agent learns which actions to take in an environment in order to maximise a cumulative reward. Through trial and error, the agent explores a wide variety of consecutive actions and evaluates them based on the rewards received. Over time, it learns a strategy for selecting appropriate sequences of actions.

Supervised learning remains the most widely used learning strategy and is also applied in this dissertation [95]. Other common variations and hybrid learning strategies include semi-supervised, self-supervised, federated, and active learning, and are further explained in Section 3.3.5.1.

3.2.3. Model architecture: Deep artificial neural networks

An ML or DL model can be seen as a parametric mapping function that is optimised to map an input to the desired output. The model architecture used in DL models is a deep artificial neural network (ANN). The structure of ANNs is inspired by the human brain, a complex network of interconnected processing units called neurons [96]. The basic building block of an ANN is an artificial neuron or perceptron, visualised in Figure 3.2. The relation between the input and output of a perceptron is given by Equation 3.1. The perceptron multiplies the inputs (x_i) with weights (w_i), sums all weighted inputs together with a bias (b), and applies an activation function (σ) to this sum to produce the output (y) [96].

A fully connected ANN is formed by placing multiple perceptrons in layers, all connected to each other such that the output of one layer becomes the input of the next layer [96]. This straightforward ANN structure is also called a multilayer perceptron (MLP). As indicated in Figure 3.3, the first and last layers are the input and output layers, respectively, while all layers in between are hidden layers. A deep ANN has at least two, and often many more, hidden layers. To model non-linear relationships between input and output, non-linear activation functions (σ) have to be introduced into the network [96]. The most common activation function is rectified linear unit (ReLU), which simply sets negative input values to zero [96, 98]. Many improvements on this activation function were introduced: Leaky

ReLU [99] introduces a small gradient for negative inputs, and Parametric ReLU (PReLU) [100] introduces a learnable parameter to define the optimal gradient. ANNs are powerful because, in theory, they can approximate almost any continuous mapping function, provided they contain at least one hidden layer and non-linear activation functions, a property described in the universal approximation theorem [101].

$$y = \sigma \left(b + \sum_{i=1}^N w_i x_i \right) \tag{3.1}$$

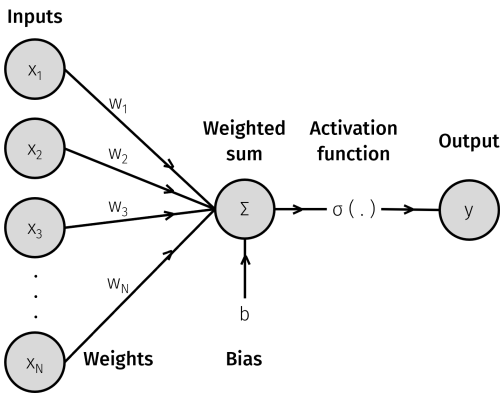


Figure 3.2. The perceptron as the basic building block of an artificial neural network.

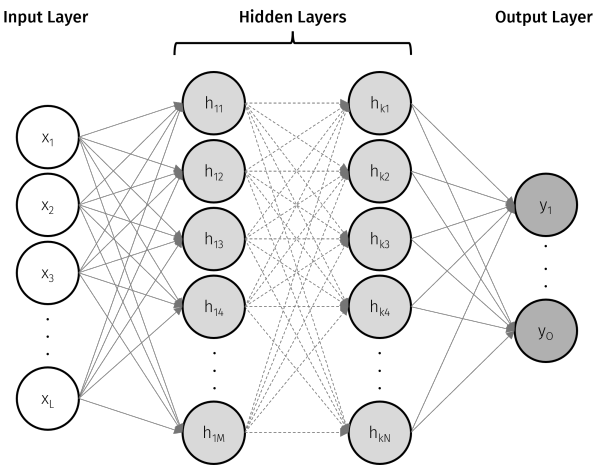


Figure 3.3. Multilayer perceptron or fully connected artificial neural network.

3.2.4. Model training

The goal of training an ANN is to find suitable values for the model parameters (weights and biases) that transform inputs into the desired outputs [96]. Model performance during training is monitored using a loss function. In supervised training, this loss quantifies the discrepancy between predicted outputs and ground-truth labels and should be minimised. The problem of finding the model parameters for which the loss function is minimised cannot be solved analytically. Therefore, the parameters of the ANN are randomly initialised and are optimised using an iterative process, as illustrated in Figure 3.4.

Each iteration consists of three steps. First, during forward propagation, the model generates predictions for all training inputs, often grouped in batches, and the loss is calculated [96]. Second, the loss is backpropagated from the output to the input layer. At every layer, the gradient of the loss with respect to the model weights is computed using the chain rule [96]. In deep networks, repeated application of the chain rule can lead to extremely small (vanishing) or large (exploding) gradients, hindering stable training. Vanishing gradients can be mitigated by choosing an appropriate activation function like ReLU which preserves non-zero gradients for positive inputs. Additional techniques to address these issues are discussed later in this chapter. Third, model parameters are updated in the direction of the negative gradient with a step size indicated by the learning rate [96]. The learning rate is typically reduced when approaching the loss minimum to prevent overshooting. This optimisation procedure is known as gradient descent (GD).

Many variations on this optimisation algorithm exist. In standard GD, the loss is calculated across the entire training dataset before updating the model parameters. One pass of all training data through the network is called one epoch. A widely used alternative is stochastic gradient descent (SGD), where parameter updates are performed after processing a single data sample or a mini-batch of data, resulting in multiple updates per epoch. Convergence can be further accelerated using adaptive learning rates for individual model parameters, indicated as adaptive stochastic gradient descent (ASGD). In addition, momentum can be incorporated, which smooths updates by using a moving average of past gradients. A common implementation is Nesterov momentum [102].

Besides the learnable model parameters (weights and biases), a large amount of hyperparameters need to be specified prior to model training [96]. These include the network architecture (number of layers and neurons per layer), activation functions, optimisation algorithm, batch size, learning rate, etc. [96]. Hyperparameter choices can strongly influence final model performance and therefore require careful fine-tuning, as further explained in Section 3.2.5.1 [96].

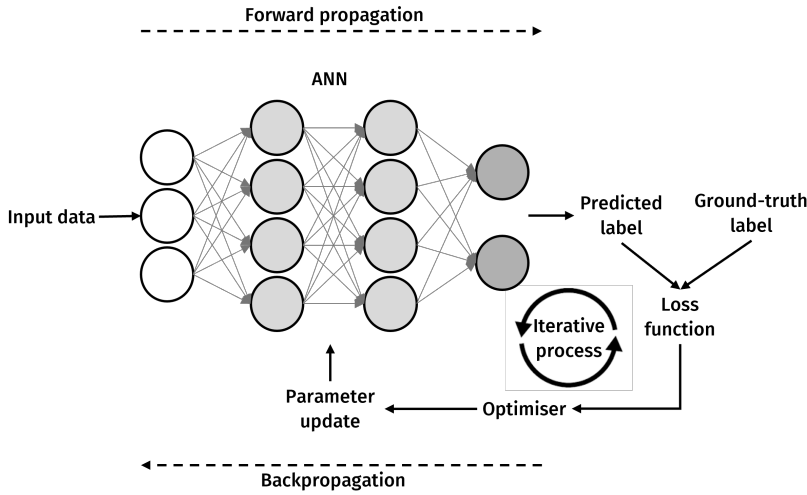


Figure 3.4. Iterative training of an artificial neural network (ANN).

3.2.5. Model generalisation and regularisation

While ANNs can model highly complex functions to accurately fit the training data, a major challenge in ML and DL is ensuring that trained models generalise well. This means that trained models should not only perform well on the training data but also on new, unseen data [96]. A model that merely memorises the training data without capturing the underlying patterns or representative features lacks generalisation capacity and is therefore ineffective when applied to previously unseen data. This section discusses the importance of appropriate data splitting to assess generalisation performance, explains the bias-variance trade-off and its relation to overfitting and underfitting, and reviews regularisation techniques aimed at improving model generalisation.

3.2.5.1. Data splits and cross-validation

To assess generalisation performance, the available dataset is typically divided into three subsets: a training, validation, and test set [96]. The training set is used to optimise the model parameters, including weights and biases. During training, model performance is monitored on the validation set to estimate generalisation performance. Although the validation set is not used to update model parameters directly, it can guide hyperparameter tuning and model selection. Grid search is a common method for selecting appropriate hyperparameters. Candidate values are defined for each hyperparameter, after which their combinations are systematically evaluated to optimise model performance. For each

configuration, the model is trained on the training set and evaluated on the validation set to select the best-performing hyperparameter combination. The search may be exhaustive or restricted to a subset of combinations using more computationally efficient search strategies. The test set remains strictly independent throughout the entire process and is used only after model training is completed to evaluate model performance on previously unseen data [96]. Ensuring independence is essential as data leakage, where information from the test set influences training or validation, can result in overly optimistic and unreliable performance estimates.

To maximise data utilisation while preventing data leakage, cross-validation schemes are often applied. In five-fold cross-validation, the data is divided into five equal partitions. In each iteration, one partition serves as the validation set, while the remaining four form the training set. This scheme allows using all available data for both training and validation. Several variations of cross-validation exist. Stratified cross-validation preserves class distributions across partitions to enable reliable model training and validation. Group cross-validation prevents leakage from related data samples, e.g. multiple images from the same patient, by keeping them within the same partition. Furthermore, when datasets are small, cross-validation can be used to estimate model performance instead of relying on a fixed held-out test set. This results in a nested cross-validation scheme, as shown in Figure 3.5, where the outer (test) folds are used to estimate the final model performance, while the inner (validation) folds are used for hyperparameter tuning.

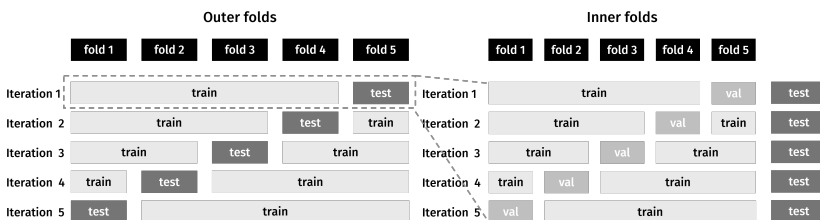


Figure 3.5. Nested five-fold cross-validation.

3.2.5.2. Bias-variance trade-off

Obtaining good generalisation starts with selecting an appropriate level of model complexity or capacity (i.e., the number of layers and neurons in the ANN) [96]. An overly simple model fails to capture the underlying structure of the data. Such a model has a high bias and underfits the training data, performing poorly on both the training and test sets. On the other hand, an overly complex model will adapt too closely to the training

data, including the noise in the data. Such model has a high variance and overfits the training data, resulting in poor performance on the test data despite near-perfect performance on the training data. This balance between underfitting and overfitting is also known as the bias-variance trade-off [103]. Figure 3.6 illustrates the bias-variance trade-off for a simple binary classification example.

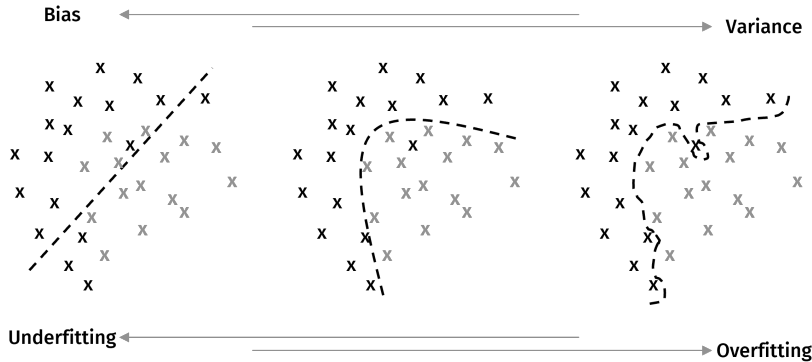


Figure 3.6. Bias-variance trade-off when choosing a decision boundary for a binary classification problem.

3.2.5.3. Regularisation techniques

In addition to selecting the right model capacity, various regularisation techniques can be applied to limit overfitting and underfitting and improve the model's generalisation performance. Several regularisation techniques that can be applied when training an ANN are discussed below [96, 104].

Data augmentation is used to artificially create new data samples based on the existing training set [96]. While large amounts of data prevent overfitting, in practice, the amount of available labelled training samples is often limited. In this case, additional data can be created by applying transformations and alterations to the existing data [104]. These transformations typically include random geometric operations (e.g., rotation and flipping) and intensity modifications (e.g., addition of Gaussian noise, Gaussian smoothing, or contrast adjustment). By exposing the model to such variations, data augmentation promotes model invariance to these alterations, increases robustness to image degradations, and leads to better generalisation performance.

Transfer learning refers to the technique in which a model trained on a particular dataset or task is reused to solve a different but related problem [96]. Rather than training a model from scratch, knowledge inferred from the initial task is transferred to the new task. Typically, the model is first trained on a large related dataset to learn general features

and patterns, after which it is fine-tuned on the target dataset, reducing the amount of task-specific data required. For example, many image analysis models are pre-trained on ImageNet, a large-scale natural image dataset introduced by Deng et al. [105] in 2009. The generic visual features learned from this dataset can be transferred to other tasks, enabling faster convergence and improved performance.

Early stopping is a regularisation strategy in which training is stopped when the validation loss reaches a minimum [96]. As illustrated in Figure 3.7, the training loss typically decreases steadily with each training epoch, whereas the validation loss may start to rise at a certain point, signalling overfitting. In early stopping, the validation loss is monitored and training is terminated when no improvement is observed for a predefined number of epochs. The model parameters corresponding to the lowest validation loss are then selected.

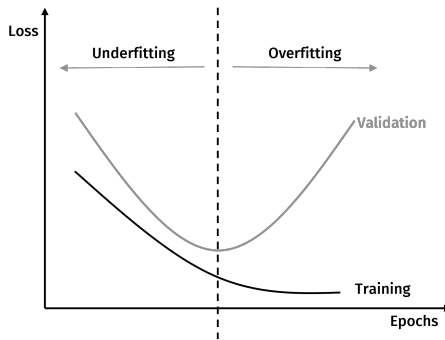


Figure 3.7. Early stopping: the dotted line indicates the number of epochs at which the model will generalise best towards unseen data.

Dropout is a technique where a fraction of neurons in the ANN is randomly deactivated during training, meaning their outputs are set to zero and they do not contribute to forward or backward propagation [96]. Dropout prevents over-reliance on specific neurons, and forces activated neurons to perform well regardless of which other neurons are included in the network. This encourages the extraction of more robust and general features, thereby improving generalisation performance. Dropout only affects the training procedure and is not used during inference.

Weight decay is a technique that discourages large model weights by adding a penalty term to the loss function [96]. Common approaches include L2 and L1 regularisation. The L2 penalty adds the sum of squared model weights to the loss function, reducing the influence of individual features by promoting smaller and more distributed weight values. The L1 penalty adds the sum of absolute model weights to the loss function, performing implicit feature selection by driving some parameters to zero. Both methods favour simpler models with improved generalisation.

3.3. Deep learning for computer vision in medical images

The application of DL in medical imaging has expanded rapidly in recent years [96]. DL techniques are being integrated throughout the entire imaging pipeline, ranging from improvements in scanner performance and image quality to automated disease detection and diagnosis [96]. In this work, we focus on the use of DL for computer vision tasks, referring to the automated analysis and interpretation of visual data. DL models can enhance the efficiency and accuracy of medical image analysis by exploiting complex image characteristics (e.g., contrast variations, shapes, and textures) that are not always readily apparent to human observers [104]. Owing to these capabilities, DL has achieved state-of-the-art performance in numerous medical computer vision tasks, particularly in image segmentation [96, 104, 106, 107]. In this work, DL models are employed to segment breast tumours in micro-PET-CT images, enabling the detection and localisation of positive resection margins.

3.3.1. Classification and segmentation

In this dissertation, we focus on two main subfields of computer vision: image classification and image segmentation. The objective of medical image **classification** is to predict diagnostic or prognostic outcomes based on imaging data. In classification tasks, one or more labels are assigned at the image level. The output of a classification model typically consists of class probabilities, which can be thresholded to determine the final class label(s). The goal of medical image **segmentation** is to accurately delineate the boundaries of anatomical or pathological structures of interest. Although different subtypes of segmentation problems exist, this work focuses exclusively on semantic segmentation. In semantic segmentation, a class label is assigned to each individual pixel in the image, thereby generating a pixel-wise annotation of the structures of interest. Throughout this dissertation, the term segmentation always refers to semantic segmentation. Segmentation models typically produce probability maps in which each pixel is associated with a probability of belonging to a certain class. These probability maps can be thresholded to obtain a final segmentation mask. In this work, we focus on binary classification and binary semantic segmentation of micro-PET-CT images of lumpectomy specimens. For margin status classification, two outcomes are considered: positive or negative margins. For breast tumour segmentation, each image pixel is assigned one of two labels: malignant or benign.

3.3.2. Convolutional neural networks

In computer vision applications, a specific type of ANN is widely used: the convolutional neural network (CNN) [96]. Although a fully connected ANN, as described in Section 3.2.3, could be applied to pixel values, this approach is highly inefficient. Medical images may contain millions of pixels. Therefore, fully connected ANNs would require a large number of parameters, resulting in excessive memory usage and computational cost [96]. Moreover, fully connected ANNs lack translational invariance [96]. Each neuron detects features at fixed spatial locations in the image, such that even small translations of the input image may lead to substantially different outputs.

CNNs address these limitations through sparse connectivity and parameter sharing [96]. Instead of fully connected neurons, CNNs contain convolutional layers with sets of learnable filters or kernels that are convolved with the input image to extract features. During a convolution operation, a kernel is systematically scanned across the image using a predefined step size or stride. At each spatial location, the sum of the element-wise product between the kernel weights and the corresponding image patch is computed. The result is a feature map that indicates the presence of the pattern encoded by the kernel at different image locations. The kernel weights form the trainable parameters of the network. Figure 3.8 illustrates such a convolution operation in 2D. By reusing the same kernel across the entire image, CNNs enforce parameter sharing and achieve translational invariance. A spatial shift in the input image will produce the same shift in the extracted feature maps. Stacking multiple convolutional layers yields a hierarchical feature representation, in which progressively more complex and abstract patterns are extracted. These learned representations can subsequently be used for tasks such as image classification or segmentation.

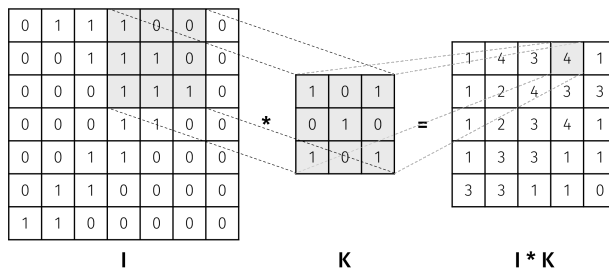


Figure 3.8. 2D convolution operation. A 7x7 input image (I) is convolved with a 3x3 kernel (K) with a stride of 1. Figure adapted from [108].

3.3.2.1. Different layer types

A CNN can have different types of layers, stacked in various configurations [96]. The most important layer types are described below.

Convolutional layers form the backbone of a CNN. They consist of several kernels, containing the trainable parameters of the network [96]. These kernels are convolved with the input to perform feature extraction. The size of each kernel (e.g., 3x3 or 5x5) defines its receptive field, and a stride > 1 can be used to reduce the input size. A convolutional layer outputs a set of feature maps, one feature map per kernel.

Fully connected layers of neurons can be applied at the end of the network, after the convolutional layers, in case of a classification problem [96]. While the convolutional layers extract different features from the input images, the fully connected layers combine these features to determine the final classification output. Such layers are not required for segmentation problems, where fully convolutional networks are applied.

Activation layers introduce non-linearities into the network through activation functions (e.g., PReLU) [96]. The final activation layer constrains the model output values to the desired range. For binary classification or segmentation problems, a sigmoid activation is used to obtain an output in $[0,1]$, representing the probability that an image or image pixel belongs to the positive class. For multiclass problems, a softmax activation is used to produce class probabilities in $[0,1]$ that sum to one across all classes. In the binary setting, sigmoid and softmax are equivalent.

Pooling layers are subsampling operations that reduce the spatial dimensions of the input by calculating summary statistics (e.g., maximum, median, or average) over local neighbourhoods [96]. These neighbourhoods commonly have a size of 2x2, thereby halving the input resolution. Alternatively, similar downsampling can be achieved by introducing strides into the convolutional layers. This dimensionality reduction decreases computational cost and enlarges the receptive field of calculated features relative to the input image. In segmentation networks, upsampling or transposed convolutional layers are also required to increase the spatial resolution of feature maps, restoring them to full size for the generation of segmentation outputs.

Normalisation layers ensure intermediate normalisation of the feature maps in the network to zero mean and unit variance [96]. The goal of normalisation layers is to increase training stability and improve convergence speed, by avoiding vanishing and exploding gradients in deep CNNs. Different types of normalisation layers exist, performing normalisation across different groups of feature maps and batch instances in one layer. The differences between instance, batch, layer, and group normalisation are illustrated in Figure 3.9.

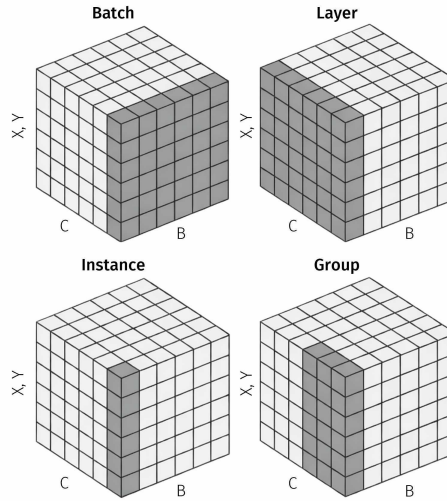


Figure 3.9. Illustration of various normalisation methods used in convolutional neural networks: batch, layer, instance, and group normalisation. A single layer of a 2D network is represented with batch (B), feature/channel (C), and spatial (X,Y) dimensions. The grey regions indicate all feature maps and batch instances that are jointly normalised for each method. In group normalisation, the number of feature maps per group can vary. Group normalisation including all channels is equivalent to layer normalisation, while group normalisation with one channel corresponds to instance normalisation. Figure adapted from [109].

Dropout layers randomly set input elements to zero during training to improve generalisation performance of the network [96].

3.3.2.2. ResNet for classification

The Residual Network (ResNet), introduced by He et al. [100] in 2016, is a widely used CNN architecture for medical image classification [96]. As illustrated in Figure 3.10, the network consists of stacked convolutional layers that progressively extract more abstract features from the input image while reducing spatial resolution. A final fully connected layer combines all extracted features to produce the classification output.

Although increasing network depth generally improves classification performance, He et al. [100] observed that beyond a certain depth, training accuracy saturates and may even degrade. This behaviour is not caused by overfitting, but is due to challenges associated with training and optimising increasingly deep CNNs. In principle, a deeper model should perform at least as well as its shallower counterpart, since the additional layers could learn an identity mapping and simply pass their input directly to the output without modification. In practice, however, learning such an identity mapping can be difficult because a CNN consists of a sequence of non-linear transformations. To address this issue, ResNet introduces residual connections, also known as shortcut connections, which bypass one or more layers in the network. These connections allow the input of a block to be directly added to its output. As a result, the network is effectively preconditioned to behave like an identity mapping, and the layers only need to learn the difference between the desired mapping and the identity mapping. This difference between the block output and its input is referred to as the residual function. In addition, the residual connections improve the flow of gradients during backpropagation, thereby mitigating the vanishing gradient problem and stabilising training. Consequently, residual connections facilitate optimisation and allow for effective training of deeper CNNs.

3.3.2.3. U-Net for segmentation

U-Net and its variants represent the most widely used CNN architectures for semantic medical image segmentation [96, 104, 110]. U-Net was originally introduced by Ronneberger et al. [111] in 2015, and was later extended to a 3D version called V-Net by Milletari et al. [112] in 2016. Today, these architectures remain state of the art for many segmentation tasks.

U-Net is a fully convolutional network that contains no fully connected layers and preserves spatial information throughout the network [111]. It maps an input image to a probability map of identical dimensions, providing pixel-level class probabilities. As illustrated in Figure 3.11, U-Net has a symmetric U-shaped architecture. The left arm is the encoder path which progressively reduces spatial resolution while extracting increasingly abstract features of the image. The encoder ends in the

bottleneck that contains a latent representation of the image. The right arm is the decoder path which subsequently upsamples the feature maps to restore the original spatial resolution while combining all extracted features to generate the final output segmentation. In addition, there are skip connections, shortcuts between corresponding encoder and decoder layers, that facilitate feature localisation and enhance model optimisation.

Numerous refinements of the U-Net architecture have been proposed [104]. To improve training stability, convergence speed, and generalisation performance, normalisation and dropout layers are commonly incorporated in the network. Furthermore, the residual variant Residual U-Net (ResU-Net) integrates residual connections within both the downsampling and upsampling blocks [113], while dense variants introduce skip connections within the blocks such that each layer gets access to all preceding feature maps within the same block [104, 110]. Beyond direct encoder-decoder skip connections, nested and densely connected skip pathways containing intermediate convolutional blocks allow fusion of features across different scales [104, 110]. Moreover, standard convolutions can be replaced by dilated convolutions, in which kernel elements are spaced apart, to enlarge the receptive field without increasing the number of parameters [104, 110].

Additional extensions include cascaded networks, stacks of encoder-decoder models that are trained end-to-end to progressively refine segmentations [104]. In these architectures, successive stages operate at increasing spatial resolutions, enabling the integration of global contextual information captured at lower resolutions with finer details extracted at higher resolutions. In some implementations, the probability map generated by one stage is provided as input to the next stage, together with the rescaled input image, to guide further refinement. Finally, for multimodal inputs, which are common in medical imaging, the different modalities can either be provided to the network as separate input channels, or processed by separate encoders that first extract modality-specific features before combining them within the network.

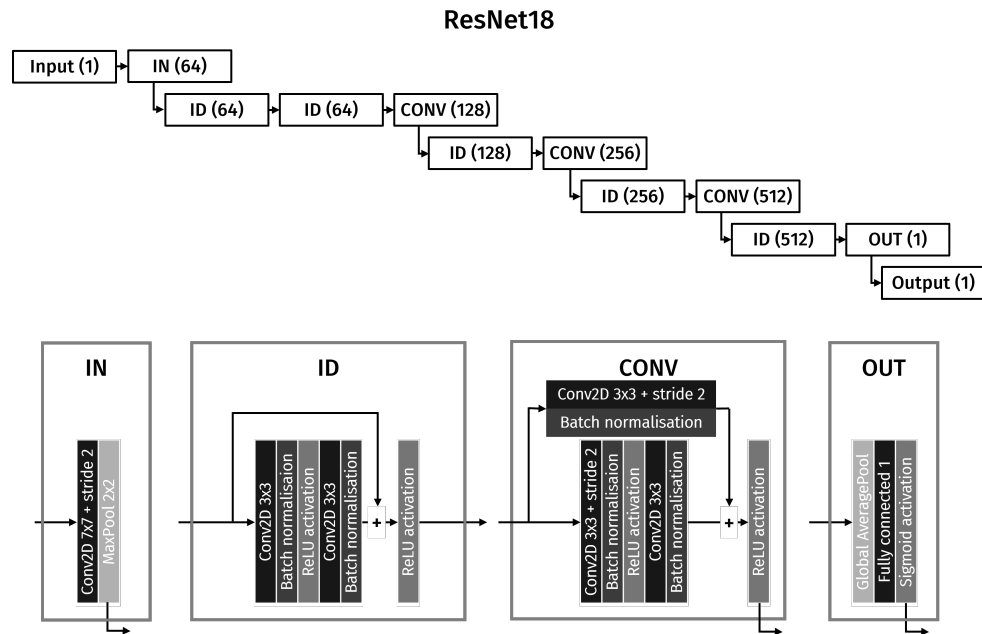


Figure 3.10. 2D Residual Network (ResNet) for image classification. The original implementation of ResNet18 according to He et al. [100] is shown for the binary classification of an input image with a single channel. The composition of the input (IN), identity (ID), convolutional (CONV), and output (OUT) units are shown. Every convolutional unit doubles the number of channels or feature maps and reduces the image dimensions by half. The fully connected layer at the end combines the extracted image features into a binary classification output.

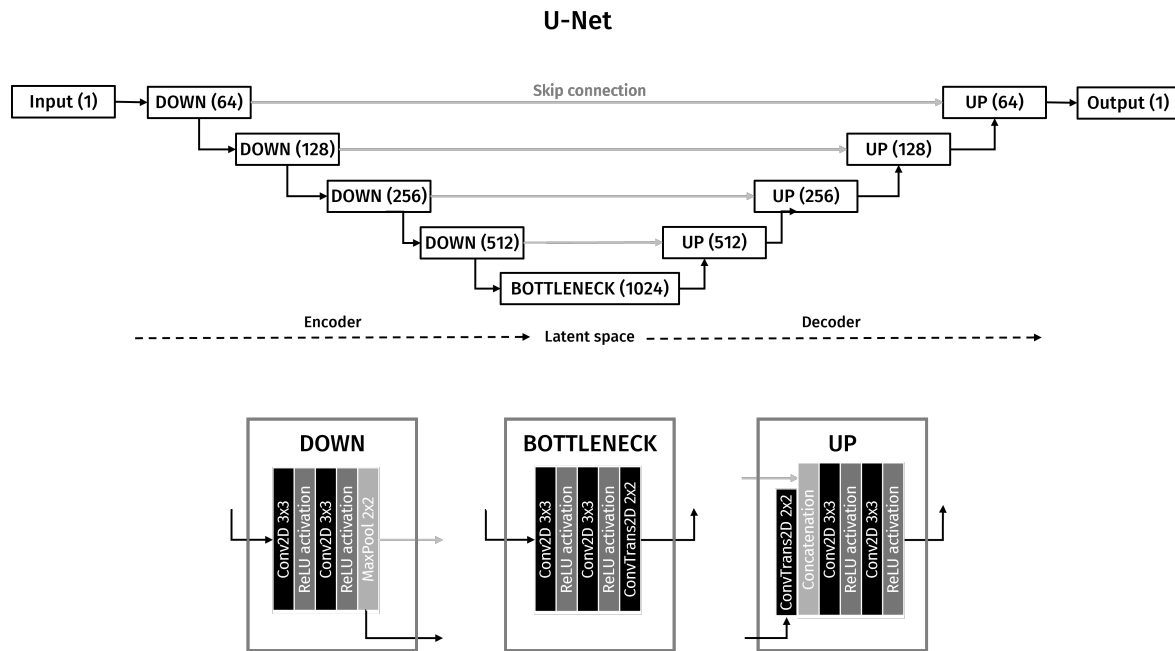


Figure 3.11. 2D U-Net architecture for image segmentation. The original implementation according to Ronneberger et al. [111] is shown for the binary segmentation of an input image with a single channel. The composition of the downsampling (DOWN), bottleneck (BOTTLENECK), and upsampling (UP) units is shown. The number of channels or feature maps is doubled in every downsampling unit and halved in every upsampling unit, while the image dimensions are halved in every downsampling unit and doubled in every upsampling unit. In the final upsampling unit, an additional convolutional layer with a 1x1 kernel is used to reduce the number of output channels to one and a sigmoid activation is applied.

3.3.3. Performance evaluation metrics

Model performance is assessed using quantitative metrics that measure the agreement between model predictions and ground-truth labels. Here, we focus on performance metrics for binary classification and segmentation tasks, although most metrics can easily be extended to multiclass problems. As medical problems are often characterised by strong class imbalance, not all evaluation metrics are equally suitable. In classification problems, the positive (e.g., disease) class often represents a minority of the population. Similarly, in segmentation problems, the region of interest (e.g., tumour tissue) typically only occupies a small fraction of the total image volume. This imbalance is also present in the problems analysed in this dissertation. For margin status classification of lumpectomy specimens, approximately 20% of cases have positive margins, as reported in Section 2.3.5 of Chapter 2. For breast tumour segmentation in lumpectomy specimens, the tumour often only covers a limited portion of the specimen, particularly in cases of invasive lobular carcinoma (ILC) and ductal carcinoma in situ (DCIS). Consequently, performance metrics and loss functions must be selected with consideration of class imbalance to ensure a reliable and clinically meaningful assessment of model performance.

3.3.3.1. Classification

In binary classification, the predicted probability of belonging to the positive class is thresholded (e.g., at 0.5) to obtain the final class label (positive or negative). Depending on the relation between the predicted and ground-truth label of a data sample, each prediction is categorised as a correct true positive (TP) or true negative (TN) prediction, or an incorrect false positive (FP) or false negative (FN) prediction. All predictions are summarised in a contingency table, also called a confusion matrix, as illustrated in Figure 3.12.

From this contingency table, several performance metrics can be derived, including accuracy (Equation 3.2), sensitivity or recall (Equation 3.3), specificity (Equation 3.4), and precision (Equation 3.5). In imbalanced classification settings where the positive class is under-represented, metrics such as accuracy and specificity may yield overly optimistic performance estimates. In these cases, sensitivity and precision provide more informative measures of performance. The F1 score (Equation 3.6), defined as the harmonic mean of recall and precision, offers a balanced measure of overall classification performance.

		Predicted Label	
		Positive	Negative
Ground-truth label	Positive	True positive (TP)	False negative (FN)
	Negative	False positive (FP)	True negative (TN)

Figure 3.12. Contingency table.

$$\text{accuracy} = \frac{TP + TN}{TP + FN + TN + FP} \quad (3.2)$$

$$\text{sensitivity (recall)} = \frac{TP}{TP + FN} \quad (3.3)$$

$$\text{specificity} = \frac{TN}{TN + FP} \quad (3.4)$$

$$\text{precision} = \frac{TP}{TP + FP} \quad (3.5)$$

$$\begin{aligned} \text{F1 score} &= \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \\ &= \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \end{aligned} \quad (3.6)$$

In practice, a trade-off exists between sensitivity, specificity, and precision, that can be controlled by adjusting the decision threshold used to convert predicted class probabilities into class labels. This trade-off can be visualised using the receiver operating characteristic (ROC) and the precision-recall (PR) curve. ROC-curves plot sensitivity or recall against one minus specificity across decision thresholds ($\in [0,1]$), whereas PR-curves plot recall against precision. Depending on the exact problem, other thresholds or operating points along these curves can be chosen. The area under the curve (AUC) ($\in [0,1]$) is a measure of the overall ability of the model to distinguish between the positive and negative class. For imbalanced problems, area under the curve of the precision-recall curve (PR-AUC) is generally more informative than area under the curve of the receiver operating characteristic (ROC-AUC). Figure 3.13 shows illustrations of the ROC- and PR-curves, including their corresponding AUC values.

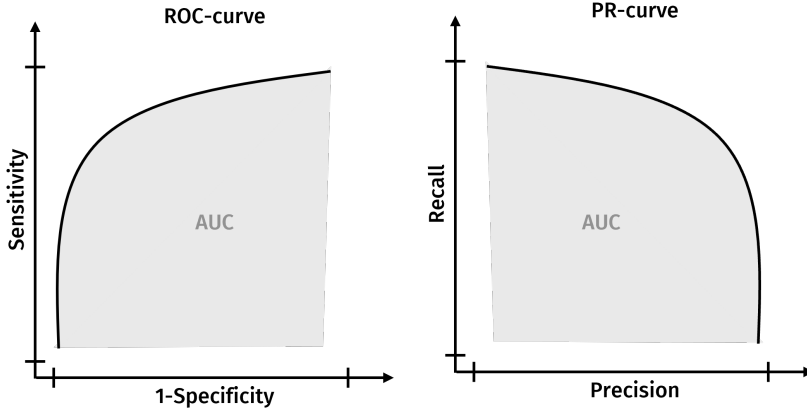


Figure 3.13. Receiver operating characteristic (ROC) and precision-recall (PR) curve. The area under the curve (AUC) is also indicated in both cases.

During training, classification models are commonly optimised using the cross-entropy loss, also called log-loss [96]. Equation 3.7 represents the binary cross-entropy (BCE) loss calculated across N data samples, with y_i the ground-truth label ($\in \{0,1\}$) and $\hat{p}_i$ the predicted probability for the positive class ($\in [0,1]$) for sample i . In the presence of class imbalance, weights can be introduced in the loss function to balance class contributions.

$$\text{BCE loss} = -\frac{1}{N} \sum_{i=1}^N [y_i \cdot \log(\hat{p}_i) + (1 - y_i) \cdot \log(1 - \hat{p}_i)] \quad (3.7)$$

3.3.3.2. Segmentation

A segmentation task can be interpreted as a pixel-wise classification problem. Consequently, after thresholding the predicted tumour probability map (e.g., at 0.5) to obtain a binary segmentation mask, pixel-level accuracy, sensitivity, specificity, precision, and F1 score can be computed. Similarly, ROC-AUC and PR-AUC can be calculated per image to assess the model's ability to discriminate between positive and negative pixels. As in classification tasks, accuracy, specificity, and ROC-AUC are less informative in imbalanced segmentation problems.

A widely used metric for segmentation is the Dice similarity coefficient (DSC), also called Sørensen-Dice coefficient [96]. The DSC ($\in [0,1]$) quantifies the spatial overlap between predicted and ground-truth segmentations, making it particularly suitable for imbalanced medical segmentation tasks [114]. Equation 3.8 shows the calculation of the DSC

for an image with N pixels, where y_i denotes the true label ($\in \{0,1\}$) and $\hat{y}_i$ denotes the thresholded prediction ($\in \{0,1\}$) for pixel i . It can be seen that the pixel-level F1 score of an image corresponds to the DSC.

$$\text{DSC} = \frac{2 \cdot Y \cap \hat{Y}}{|Y| + |\hat{Y}|} = \frac{2 \cdot \sum_{i=1}^N y_i \cdot \hat{y}_i}{\sum_{i=1}^N y_i + \sum_{i=1}^N \hat{y}_i} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}} \quad (3.8)$$

In addition to overlap-based metrics, metrics focusing on the predicted segmentation contour or boundary are also commonly used. The Hausdorff distance quantifies the maximum distance one can be forced to travel from a point on the predicted contour to the ground-truth contour, or the other way around. To reduce the effect of outliers, percentile variants (e.g., 95th percentile of the Hausdorff distance (95HD)) are commonly applied. Contour Dice metrics can likewise be computed to assess the overlap between the predicted and ground-truth contours. The score typically incorporates a spatial tolerance (e.g., contour Dice with a 1 mm tolerance (1mmCD)) to allow for minor boundary deviations.

Given the widespread use of the DSC as an evaluation metric, the Dice loss is commonly employed during training of segmentation models [115]. Equation 3.9 shows the calculation of this loss for an image with N pixels, where y_i is the true label ($\in \{0,1\}$) and $\hat{p}_i$ is the predicted tumour probability ($\in [0,1]$) for pixel i .

$$\text{Dice Loss} = 1 - \frac{2 \cdot \sum_{i=1}^N y_i \cdot \hat{p}_i}{\sum_{i=1}^N y_i + \sum_{i=1}^N \hat{p}_i} \quad (3.9)$$

3.3.4. Challenges related to deep learning for medical image analysis

DL-based medical image analysis comes with its own particular challenges related to medical imaging data. Below, we discuss the most important challenges.

3.3.4.1. Scarcity of well-curated annotated datasets

A major challenge in developing DL models for medical imaging applications is the limited availability of large, well-curated datasets [96, 110]. Access to medical data is strictly regulated to protect patient privacy (e.g., through GDPR in Europe and HIPAA in the US). Data sharing requires patient consent, proper data anonymisation, and secure data handling at all times. These legal and administrative requirements introduce logistical challenges that complicate access to large-scale datasets.

In addition, supervised learning relies on ground-truth annotations, which typically require manual image classification or segmentation by medical experts. This process is labour-intensive and time-consuming [96]. Furthermore, annotations are subject to intra- and inter-reader variability, introducing uncertainty and potential inconsistencies in the labels [104].

3.3.4.2. Class imbalance

As previously mentioned, many medical imaging tasks are characterised by class imbalance [96, 110]. In the case of classification tasks, the positive (e.g., disease) class is often under-represented, whereas in segmentation problems, the region of interest (e.g., tumour) typically occupies only a small fraction of the image volume. To prevent overly optimistic performance estimates, balanced loss functions and appropriate evaluation metrics should be employed during model training and assessment.

3.3.4.3. High-dimensional complex data

Medical image analysis is inherently challenging due to the high-dimensional and complex nature of the data. Medical imaging datasets are often 3D and may involve multimodal information, e.g. combining functional and anatomical signals [96]. This increases both computational demands and modelling complexity. Furthermore, data is often of low quality, containing noise, artefacts, or low contrast. These issues are particularly problematic in segmentation tasks, where target structures often have complex irregular shapes and weak or low-contrast boundaries relative to surrounding tissues [104, 110].

3.3.4.4. Data heterogeneity and domain shifts

Beyond data complexity, medical image analysis is also challenging due to intrinsic and extrinsic sources of data variability. Intrinsic heterogeneity arises from substantial inter-patient variability [96, 110], as both normal anatomy and pathological structures can differ considerably in shape, intensity, and texture. DL models must therefore be sufficiently robust to generalise across diverse patient populations. Extrinsic variability stems from differences in scanner hardware and acquisition protocols across institutions, affecting image resolution, contrast, intensity distributions, noise characteristics, and artefacts [96, 104, 110]. While spatial and intensity standardisation can be applied prior to model training, considerable variability remains.

When such intrinsic or extrinsic differences lead to discrepancies between training and deployment data distributions, commonly referred to as domain shifts, model performance may degrade substantially. To mitigate these effects and avoid bias, models should be trained and validated on large, diverse, multi-institutional datasets that capture both biological and acquisition-related variability to ensure robust generalisation across hospitals, scanners, and patient populations.

3.3.5. Emerging trends in deep learning for medical image analysis

3.3.5.1. Data-Efficient and collaborative learning strategies

Synthetic data generation is increasingly used to expand training datasets due to the limited availability of annotated medical imaging data. Generative adversarial networks (GANs) are widely used for this purpose [104]. A GAN consists of a generator that maps random noise to synthetic images, and a discriminator that distinguishes real from generated images. Through adversarial training, a process where both networks compete with each other, the generator learns to produce increasingly realistic medical images and eventually fools the discriminator. More recently, diffusion models have emerged as a powerful alternative to GANs. Diffusion models learn to progressively transform random noise into realistic images, often providing more stable training and higher-quality results.

Semi-supervised learning is a hybrid learning approach in which a model is trained on a dataset containing a small proportion of labelled data alongside a larger amount of unlabelled data [104]. A common strategy involves first training the model on the labelled subset and subsequently predicting pseudo-labels for the unlabelled samples. Predictions with high confidence are then gradually incorporated into the training set, and the model is retrained to iteratively refine these pseudo-labels.

Self-supervised learning trains models using data for which labels are automatically generated from the data itself. A so-called pretext task is used to train the model, encouraging the model to learn meaningful and transferable feature representations of the data [104]. Self-supervised learning is often used as a pre-training step, after which the model is fine-tuned on a (limited) labelled dataset to perform a specific downstream task. A common type of pretext task is contrastive learning [104]. In this technique, positive and negative image pairs are constructed, and the model is trained using a loss function that promotes similar feature representations for positive pairs and dissimilar representations for negative pairs. Positive pairs may for example consist of different augmented versions of the same image. Other pretext tasks, specific for image segmentation, include rotation prediction, jigsaw puzzle solving,

and image inpainting. Self-supervised learning is also commonly used to develop foundation models, large-scale models that are pre-trained on diverse datasets to learn general-purpose representations transferable across tasks and domains. An example is MedSAM (“Segment Anything in Medical Images”), a foundation model trained to perform segmentation tasks in a broad range of medical images [116].

Federated learning enables the distributed training of DL models without requiring the exchange of data between institutions, thereby addressing concerns related to sharing personal health information [104, 110]. In this setup, a central server hosts a global model, which participating hospitals download and train locally on their own data. The locally updated model parameters are then sent back to the central server, where they are aggregated to update the global model. This approach facilitates safe multicentre collaboration while preserving patient privacy.

Active learning is a learning strategy that promotes efficient data labelling. A model is initially trained on a small labelled dataset and subsequently used to iteratively identify additional data samples that would be most informative if labelled [104]. Based on measures reflecting prediction uncertainty or sample representativeness, the model selects data points for expert annotation. These newly labelled samples are then added to the training set, enabling the model to improve efficiently while requiring less annotation data.

3.3.5.2. Architectural innovations

Multi-task learning refers to training a single model to simultaneously perform two or more related tasks [104]. By learning these tasks jointly, the model can develop a shared and more robust latent representation of the input image, which often leads to improved generalisation. A single network can therefore handle multiple segmentation and/or classification tasks at the same time. In practice, some implementations use task-specific subnetworks or combine a shared encoder with separate task-specific decoders. This design enables partial parameter sharing across tasks, while still allowing each task to be fine-tuned individually using its own dedicated loss function.

Attention mechanisms introduce adaptive feature weighting into CNNs. Rather than treating all extracted image features equally, attention selectively focuses on task-relevant information, while suppressing irrelevant background signals [104]. Attention modules are typically lightweight modules that operate either at the channel level (indicating which features are relevant) or at the spatial level (indicating which regions to focus on). In U-Nets, attention gates are commonly integrated into the skip connections [104]. Importantly, attention mechanisms enhance the

representational capacity of CNNs without fundamentally altering their convolutional structure.

Vision transformers adopt a fundamentally different network architecture. While CNNs rely on convolutional kernels with local receptive fields, transformers use self-attention mechanisms to model pairwise interactions between image patches [104, 110]. This enables global context modelling within a single layer and facilitates the capture of long-range anatomical dependencies. However, compared to their CNN counterparts, vision transformers generally require larger training datasets and are computationally expensive, particularly for high-resolution images. Therefore, in medical image analysis, hybrid CNN-transformer architectures have become popular, combining strong local feature extraction with global context modelling through self-attention.

Self-configuring networks aim to reduce the need for manual hyperparameter tuning and expert-driven design choices in DL workflows. As previously mentioned, hyperparameter selection can strongly influence final model performance [96]. Medical imaging datasets differ considerably in modality, image size, spatial resolution, and intended application, often requiring tailored solutions. However, hyperparameter tuning is complex, as numerous design choices are interdependent, and overall performance will largely depend on the expertise of the researcher [104]. For semantic segmentation tasks, Isensee et al. [117] introduced nnU-Net (“No New Net”), a self-configuring framework that automatically analyses the training data and adapts a U-Net architecture to the task at hand. In addition to selecting the network architecture, nnU-Net defines a complete, dataset-specific pipeline including preprocessing, training, inference, and postprocessing. This automated approach reduces the need for extensive expert intervention while maintaining robust performance across diverse segmentation applications.

3.3.5.3. Advanced optimisation and supervision techniques

Deep supervision improves model optimisation by introducing auxiliary output branches at intermediate network layers, allowing additional loss terms to be computed from intermediate predictions. The total loss function is typically formulated as a weighted sum of the final and auxiliary losses [104]. As the loss function drives model optimisation, its design has an important influence on model performance. Deep supervision improves gradient flow and facilitates convergence, while encouraging the learning of meaningful intermediate data representations. In segmentation, deep supervision is commonly applied in the decoder branch, where intermediate feature maps are upsampled to the original image resolution and compared against the ground truth to compute auxiliary loss functions. In some architectures, these upsampled intermediate feature

maps from different decoder levels are concatenated to produce the final segmentation output, potentially improving segmentation performance by combining features at different resolutions.

Learnable loss functions can be used in image segmentation problems to go beyond pixel-wise accuracy by promoting structural and anatomical consistency. Conventional loss functions, such as cross-entropy or Dice loss, quantify pixel-wise overlap but may fail to capture higher-order structural similarities between predicted and true segmentation masks. Learnable loss functions can help address this limitation. One approach is to use a conditional GAN, consisting of a generator and discriminator trained adversarially [104]. In this case, the generator, often implemented as a U-Net, predicts a segmentation mask from an input image, while the discriminator learns to distinguish between ground-truth and generated segmentation masks. The generator is optimised using a combination of a standard segmentation loss and an adversarial loss derived from the discriminator, which encourages structurally and anatomically realistic predictions rather than merely optimising pixel-wise similarities. Auto-encoders provide another strategy to construct expressive loss functions [104]. An auto-encoder is an encoder-decoder network trained to reproduce its input. The network can be trained on ground-truth segmentation masks to learn compact and meaningful latent representations that capture their essential structural characteristics. In this way, the auto-encoder embeds prior knowledge about plausible segmentation shapes into its latent space. During training of a segmentation model, both predicted and ground-truth masks can be projected into the latent space of the pre-trained auto-encoder, and their difference can be used as a loss term to capture structural discrepancies and encourage predictions that resemble learned anatomical priors.

3.3.5.4. Techniques for efficient and trustworthy deployment

Knowledge distillation compresses the knowledge of a large model or model ensemble into a single, more lightweight model [104]. Lightweight models are essential for deploying DL systems in time-sensitive clinical environments with limited computational resources. CNNs used for computer vision tasks are often large and highly parametrised, making inference computationally expensive and time-consuming [110]. Knowledge distillation transfers the knowledge from a well-trained complex teacher network to a smaller and more efficient student network. To facilitate this transfer, the student network's loss function is extended with additional terms that penalise discrepancies between the student and teacher outputs, as well as differences in their intermediate feature representations.

Uncertainty quantification aims to improve model trustworthiness by providing quantitative confidence estimates intended to reflect the likely correctness of predictions and to help detect potential prediction errors [104]. Such transparency is essential for clinical adoption of DL models, as safe integration into practice requires understanding when and why a model may fail. Ideally, uncertainty estimates should be well calibrated, meaning that the predicted confidence accurately reflects the true probability of a correct prediction. Uncertainty generally arises from two main sources: epistemic uncertainty, which is related to the model itself, and aleatoric uncertainty, which stems from the data [104, 118, 119]. Epistemic or model uncertainty arises from a lack of knowledge about the system, and can often be reduced by collecting more complete training datasets or improving model design. A common way to estimate model uncertainty is by comparing predictions from an ensemble of models trained under different conditions. Aleatoric or data uncertainty stems from inherent noise or randomness in the data, and is often irreducible. It can be assessed using techniques such as test-time augmentation, where multiple predictions are obtained from simple transformations (e.g., flips or rotations) of the same input image.

3.4. Ground-truth generation through medical image co-registration

As previously mentioned, supervised training of DL models relies on labelled medical images. While ground-truth annotations are typically provided by medical experts, they are subject to intra- and inter-reader variability. This challenge is particularly relevant to this dissertation, where ground-truth breast tumour segmentations in micro-PET-CT images of lumpectomy specimens are required. Producing clinically meaningful contours is difficult in this case, as nuclear medicine physicians are often unfamiliar with the high resolution of micro-PET-CT images [104]. To overcome this limitation, we perform co-registration of micro-PET-CT images with corresponding histopathology slides of the same specimen and transfer tumour annotations from histology to the micro-PET-CT images. Although the specific implementation of this co-registration pipeline is described in detail in the next chapter, the present section introduces the fundamental concepts underlying classical image co-registration to facilitate understanding of the label-generation process.

Image co-registration aims to spatially align two images such that corresponding anatomical structures coincide [96]. Both images are brought into a shared reference frame in which corresponding coordinates represent the same physical region. As explained below, classical automatic registration methods formulate this task as an iterative optimisation problem. For comprehensive reviews and recent advances in image

co-registration, including DL-based approaches for similarity estimation and direct transformation prediction, we refer the reader to the work of Decuyper et al. [96], Haskins et al. [120], and Darzi et al. [121].

Classical co-registration methods follow an iterative optimisation process to estimate the parameters of a spatial transformation that aligns anatomical structures in a moving image with their corresponding structures in a fixed image [121]. This iterative process, illustrated in Figure 3.14, involves several steps: (1) the current estimate of the spatial transformation is applied to the moving image in order to align it with the fixed image, (2) the transformed moving image is interpolated to the grid of the fixed image, (3) a loss function is calculated to quantify alignment quality, and (4) an optimiser updates the transformation parameters in the direction of the negative gradient of the loss function in order to progressively improve image alignment.

The first component of this iterative framework is the spatial transformation applied to the moving image. This transformation must be parametrised [121]. In increasing order of complexity and degrees of freedom, common models are rigid, affine, and deformable transformations. Rigid transformations include translation and rotation and are typically sufficient for rigid structures such as bone. Affine transformations extend this model with scaling and shearing. Deformable or free-form transformations allow spatially varying deformations and are generally required for soft tissues that can locally deform. A common deformable parametrisation is the B-Spline model, in which a grid of control points defines smooth local deformations. For complex alignment tasks, a hierarchical approach is often adopted, progressively increasing transformation complexity and initialising each co-registration stage with the result of the previous one. The transformation parameters are iteratively optimised during co-registration. Once the moving image has been transformed, it must be resampled onto the grid of the fixed image in order to enable pixel-wise comparison between both images. Choosing an interpolation method involves a trade-off between accuracy and computational cost. Common approaches, in order of increasing complexity, are nearest neighbour, linear, and B-spline (cubic) interpolation. In practice, lower-order interpolation is often used during iterative optimisation, while a higher-order interpolation is applied to obtain the final result.

After transformation and interpolation, the alignment between the moving and fixed image must be quantified in order to guide the optimisation process [121]. Image alignment can be expressed using feature-based or intensity-based approaches. Feature-based methods minimise distances between corresponding landmarks in both images. Landmarks are often manually indicated by experts, making this approach labour-intensive, time-consuming, and error-prone. In some cases, landmarks can be detected automatically or artificial markers may be introduced prior to

imaging, although the latter can damage the tissue. Moreover, feature-based methods are generally less suitable for deformable co-registration. In contrast, intensity-based methods automatically compute similarity metrics based on the relation between the intensity distributions of both images. Several similarity metrics have been proposed for this purpose. The mean squared difference metric assumes a one-to-one relation, and is mainly used for monomodal intra-patient co-registration. The normalised correlation coefficient assumes a linear relation, and is mainly used for inter- or intra-patient co-registration of closely related modalities. Normalised mutual information assumes a statistical relation, and can be used for multimodal co-registration. These metrics can serve both as loss functions to drive co-registration optimisation and as evaluation measures for the final alignment. Finally, following the direction of the gradient of the loss function, an optimiser updates the transformation parameters to improve image alignment. To this end, gradient descent (GD) and its variants, introduced in Section 3.2.4, are commonly employed.

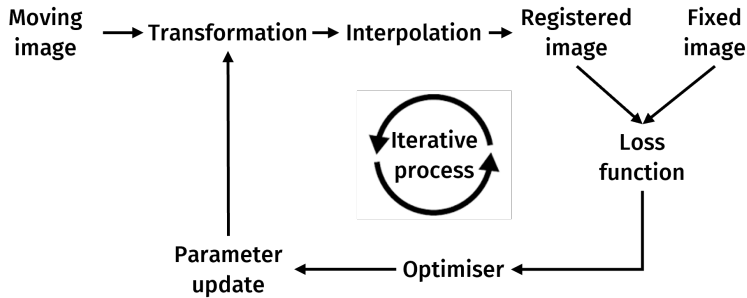


Figure 3.14. Iterative image co-registration.

Part II

Research Results

4 | Co-registration of micro-PET-CT with histopathology images of lumpectomy specimens to obtain ground-truth tumour segmentations

Abstract

Introduction: To correctly interpret positron emission tomography (PET) images at submillimetre resolution, it is important to gain insight into radiotracer distributions in tissues at high resolution. To link the molecular imaging signals to their histological origins, we propose a method for co-registration of high-resolution (micro-) *ex vivo* specimen PET images, combined with computed tomography (CT) images, and the corresponding specimen histopathology. We apply this method to breast cancer (BCa) specimens to collect a dataset of micro-PET-CT images with ground-truth breast tumour annotations, required for the development of deep learning (DL) models for segmentation.

Methods: The proposed co-registration method has two components. First, we use an image acquisition scheme that minimises and tracks tissue deformation: (1) We acquire micro-PET-CT images of approximately 2 mm thick lamellas of fresh surgical specimens, enclosed in tissue cassettes. (2) We acquire micro-CT images of the same lamellas after formalin fixation to visualise tissue deformation. (3) We obtain one haematoxylin and eosin (H&E) stained histopathology section per lamella of which we capture a digital whole slide image (WSI). Second, we apply an automatic co-registration algorithm to improve the alignment between the micro-PET-CT images of the fresh lamellas and WSIs, guided by the micro-CT of the fixated lamellas. We applied the co-registration method to BCa specimens of patients who were preoperatively injected with 0.8 MBq/kg [^{18}F]fluorodeoxyglucose ([^{18}F]FDG). To estimate the spatial co-registration error, we calculated the distance between corresponding microcalcifications in the micro-CTs and WSIs. The obtained co-registered images allowed us to study standardised uptake values (SUVs) of different breast tissues, as identified on the WSIs by an experienced breast pathologist.

Results: We imaged 28 BCa specimens, 19 cases of invasive carcinoma of no special type (NST), 6 of invasive lobular carcinoma (ILC), and 3 of ductal carcinoma in situ (DCIS). In total, we applied the co-registration method to 53, 19, and 10 cassettes of NST, ILC, and DCIS specimens, respectively. While the cassette framework minimised tissue deformation, the best alignment between the micro-PET-CT images and WSIs was

achieved after deformable co-registration. We found an overall average co-registration error of 0.69 ± 0.23 mm between the micro-PET images and WSIs. (Pre)malignant tissue (including NST, ILC, and DCIS) generally showed higher SUVs than healthy tissue (including healthy glandular, connective, and adipose tissue). As expected, inflamed tissue and skin also showed high uptake.

Conclusions: We co-registered micro-PET-CT images of lumpectomy specimens with their corresponding histopathology WSIs, achieving an accuracy comparable to the spatial resolution of the micro-PET images. The aligned data provided important insights into the uptake of [^{18}F]FDG in different breast tissue types at high resolution. Furthermore, the obtained dataset of micro-PET-CT images with ground-truth breast tumour segmentations is crucial for the development and validation of DL-based segmentation models. While currently we only applied this method to BCa specimens, we believe the proposed method is applicable to a wide range of specimens and radiotracers, providing insight into distributions of (new) radiotracers in human malignancies at a submillimetre resolution.

This chapter contains information previously published in:

Luna Maris, Menekse Göker, Jens M. Debacker, Kathia De Man, Bliede Van den Broeck, Jo Van Dorpe, Koen Van de Vijver, Vincent Keereman, and Christian Vanhove. Method for co-registration of high-resolution specimen PET-CT with histopathology to improve insight into radiotracer distributions. *European Journal of Nuclear Medicine and Molecular Imaging Physics*, 11(1):85, 2024.

Luna Maris, Vincent Keereman, and Christian Vanhove. Method for co-registration of nuclear images with microscopic histopathology images, Patent with International Application Number PCT/EP2025/054989, Publication Number WO/2025/186049, and Priority Date 08.03.2024.

4.1. Introduction

Positron emission tomography (PET), in combination with computed tomography (CT), has become widely adopted in oncology [80, 81]. In PET, different types of radiotracers can be used to visualise various tissue characteristics. In order to correctly interpret PET images, we need to understand how radiotracers distribute in different types of tissue. As the correlation of PET images with histopathology is complex, distribution mechanisms are still unclear, even for [^{18}F]fluorodeoxyglucose ([^{18}F]FDG), one of the most common radiotracers in oncology [122]. Especially for submillimetre (micro-)PET images, where radiotracer distributions are imaged at unprecedented resolutions, we need to improve our insight into the distribution of existing and new radiotracers in both cancerous and healthy tissue.

In order to study radiotracer distributions in different tissues, an accurate correlation of PET-CT images with histopathology is required. Published methods perform co-registration between *in vivo* PET-CT images and histopathology slices of the tissue after surgical resection. Examples are the methods introduced by Schiller et al. [123] and Puri et al. [124, 125] to perform co-registration for the prostate and head and neck region, respectively. Existing methods have several limitations [126]. First, most methods need extensive tissue processing to achieve a match of orientation and cut between PET-CT images and histopathology [126]. For example, Schiller et al. [123] used special embedding and cutting equipment to obtain parallel, contiguously cut, and whole-slice histopathology images of the prostate. This limitation was previously addressed by Puri et al. [124, 125], who developed a method to co-register *in vivo* PET-CT images of the head and neck region with histopathology slices obtained in routine pathology settings. Second, previously published mathematical co-registration algorithms often rely on minimising the distance between hand-picked pairs of landmarks in the PET-CT and histopathology images [124, 126]. This method is prone to error and sampling bias [127]. Other groups make use of fiducial markers (e.g., laser burns or ink spots) which lead to tissue damage and information loss [124, 128, 129]. Due to these challenges, existing co-registration methods remain time-consuming and labour-intensive, resulting in small patient cohorts [123, 124, 126]. Additionally, almost all publications focus on firm tissues that retain their shape after resection, such as prostate or laryngeal tissue [123, 124, 126]. Correlation is more challenging for soft tissue, such as breast tissue, which is prone to deformation during histopathological processing.

To overcome these challenges, we propose a novel strategy to co-register high-resolution *ex vivo* PET-CT images with histopathology whole slide images (WSIs). We use a dedicated specimen scanner to capture micro-PET-CT images of resected tissue. Acquiring *ex vivo* PET-CT images

enables maintaining control over tissue deformation, making correlation with histopathology more straightforward. Another advantage of *ex vivo* tissue scanning is the opportunity to achieve a higher spatial resolution compared to *in vivo* scanning, allowing us to study radiotracer uptake in tumour and healthy tissue at a submillimetre resolution. Several groups already showed that it is feasible to acquire *ex vivo* (micro-)PET-CT images of oncological specimens in skin [130], breast [30, 131], ovarian [132], prostate [133, 134], head and neck [87, 135], thyroid [136], pancreatic [137], colon [61, 138], liver [61], and lung [139, 140] cancer. While several groups correlated the *ex vivo* PET-CT imaging results with histopathology, an accurate and systematic approach for co-registration is lacking [30, 87, 133, 135, 138, 139]. In previous work by Debacker et al., we already performed co-registration between autoradiographs of resected human tissue and the corresponding histopathology [87]. However, we did not succeed in directly linking *ex vivo* micro-PET images with histopathology due to tissue degradation and repositioning caused by the process required to fixate and slice the tissue for histopathological analysis. In this work, we address these limitations and develop a co-registration method that minimises, tracks, and compensates for tissue deformation. Our method to accurately align micro-PET-CT images of resected tissue with the corresponding histopathology is inspired by the work of de Boer et al. [127], who performed co-registration of optical measurements of breast tissue with histopathology.

As an example application, we evaluate the proposed co-registration method in breast cancer (BCa). We align [^{18}F]FDG micro-PET-CT images of resected BCa specimens with the corresponding histopathology. These co-registered images can then be used to study the standardised uptake values (SUVs) of different types of breast tumours at a submillimetre resolution. PET currently has a limited role in BCa diagnosis as its sensitivity depends on the lesion's size and histology [88, 141–143]. It is especially difficult to detect cases of ductal carcinoma in situ (DCIS) and invasive lobular carcinoma (ILC), as these breast tumours generally show lower uptake compared to a case of invasive carcinoma of no special type (NST) [88, 141–143]. By improving our understanding of radiotracer distributions in breast lesions at a submillimetre resolution, we gain insight into whether improvement in the resolution of PET scanners will enhance the visualisation of BCa in the future. In addition, the obtained dataset of micro-PET-CT images with ground-truth breast tumour annotations is crucial for the development and validation of deep learning (DL) models for segmentation.

4.2. Methods

4.2.1. Co-registration method

The proposed co-registration method consists of two components. First, a strict scheme is followed to acquire micro-PET-CT images of resected oncological specimens, such that, later on, these micro-PET-CT images can easily be correlated with the corresponding histopathology sections. Second, to improve the alignment of both imaging modalities, an automatic mathematical co-registration algorithm is developed and applied.

4.2.1.1. Image acquisition scheme

We propose an image acquisition scheme to obtain micro-PET-CT images of surgically resected specimens of patients preoperatively injected with a PET radiotracer. This scheme, that minimises and tracks tissue deformation, consists of different steps: (1) After surgical resection, a pathologist slices the obtained specimen into lamellas of approximately 2 mm thickness. (2) These lamellas are enclosed in standard meshed tissue cassettes. (3) Micro-PET-CT images of the cassettes containing the fresh tissue lamellas are acquired. (4) The tissue is fixated in a buffered formalin solution for at least 12 hours. (5) Micro-CT images of the fixated lamellas are acquired to visualise the tissue deformation caused by the fixation process. (6) The tissue is embedded in paraffin. (7) An approximately 5 μm thin tissue section is cut from each paraffin block, parallel to the ground plane of the cassette. These tissue sections are mounted on glass slides and coloured using haematoxylin and eosin (H&E) staining. (8) Finally, digital WSIs are obtained from these pathology slides using a slide scanner. All required tissue processing steps are part of the standard-of-care procedure for most tumour specimens. The only adaptation is that the surgical specimen needs to be sliced while it is still fresh, which is not standard practice in all pathologies. Figure 4.1 visualises this image acquisition scheme in case of a BCa specimen. However, the proposed method can be applied to any surgically resected cancer specimen.

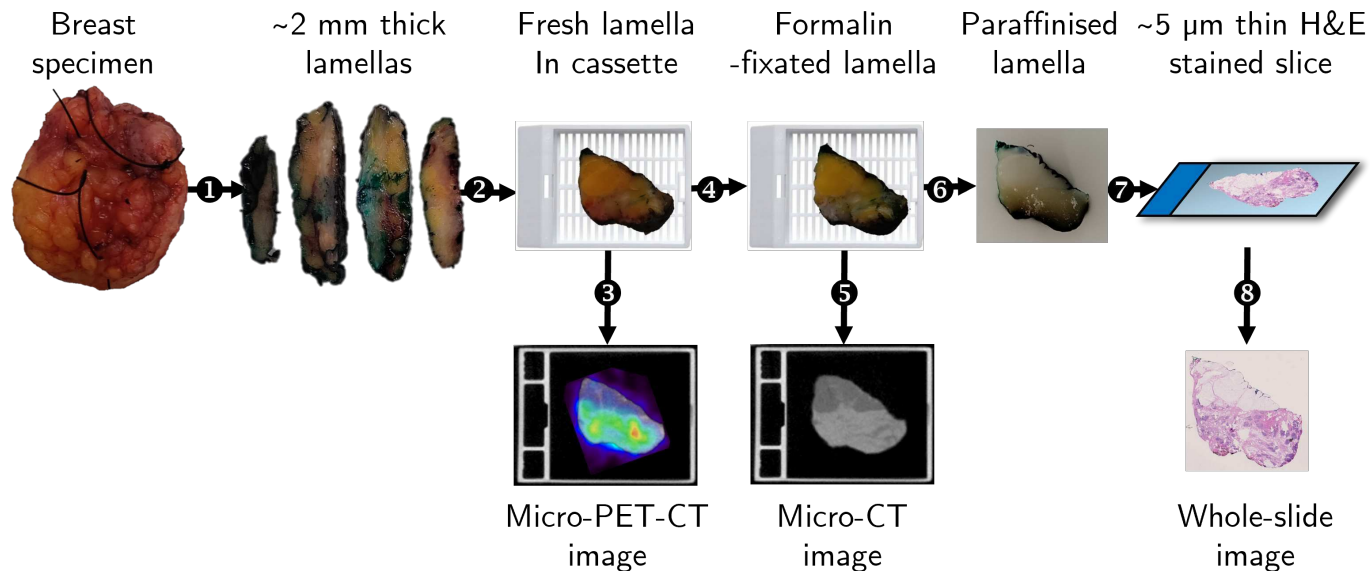


Figure 4.1. Illustration of the proposed image acquisition scheme for a breast cancer specimen. (1) Specimen is sliced into lamellas of approximately 2 mm thickness. (2) Fresh lamellas are enclosed in tissue cassettes and (3) micro-PET-CT images of the fresh tissue are acquired. (4) Lamellas are fixated in formalin and (5) micro-CT images of the fixated tissue are acquired. (6) Lamellas are embedded in paraffin. (7) From these paraffin blocks, thin tissue sections of approximately 5 µm thickness are cut, mounted on glass slides, and coloured using a haematoxylin and eosin (H&E) staining, and (8) digital whole slide images are captured from these pathology slides.

4.2.1.2. Mathematical co-registration algorithm

While the acquisition scheme minimises tissue deformation, tissue fixation and sectioning will still cause local deformations. To compensate for this, we perform iterative mathematical co-registration of the micro-PET-CT images of the fresh lamellas and the corresponding WSIs. The micro-CT images of the fixated lamellas are used to guide this algorithm. As illustrated in Figure 4.2, we apply the following four-step co-registration algorithm:

I. Co-register the micro-CTs of the fresh and fixated lamellas

First, we compensate for tissue deformation caused by fixation. We define the transformation that aligns the 3D micro-CT of a fresh lamella (moving image) with the 3D micro-CT of the same lamella after fixation (fixed image). For this step, a free-form B-spline deformation model is used to account for possibly non-uniform deformation due to differences in tissue composition. As tissue densities slightly change during fixation, the normalised correlation coefficient (NCC) is a suitable metric to match both micro-CT images.

II. Select the micro-CT slice of the fresh lamella that matches the WSI

As a second step, we select one slice of the transformed micro-CT image from step I that matches the WSI. The micro-CT is a 3D image of a tissue lamella, while the WSI is a 2D image of a thin section cut from that lamella after paraffin embedding. We match the WSI with a slice of the transformed micro-CT resulting from step I, assuming that most tissue deformation occurs during fixation, not paraffin embedding. We treat the micro-CT as a stack of 2D slices parallel to the cassette surface, with a slice thickness equal to the voxel size of the micro-CT. For every micro-CT slice, we perform an affine co-registration with the WSI, with the normalised mutual information (NMI) as metric. To express the match of the tissue compositions and shapes in both images, the sum of the NMI and the Dice similarity coefficient (DSC) is calculated. The micro-CT slice with the highest score after affine co-registration is selected.

III. Co-register the selected micro-CT slice and the WSI

As a third step, the selected micro-CT slice (moving image) is co-registered with the WSI (fixed image). As recommended by Albers et al. [144] and de Boer et al. [127], a free-form B-spline deformation is used to compensate for (possibly non-uniform) in-plane deformations caused by sectioning of the paraffin blocks and staining [144]. The NMI metric is again used to match both modalities.

IV. Apply the micro-CT transformations to the micro-PET image

Finally, the micro-PET image of the fresh lamella is aligned with the WSI, by applying the transformations found for the corresponding co-acquired micro-CT image of the fresh lamella. These transformations are derived solely from the micro-CT image, while the micro-PET signal is deliberately not used to drive the co-registration. This is important to keep the alignment independent of the tracer uptake distribution, since optimising the alignment on the PET signal itself would bias the subsequent analysis of tracer uptake across different tissue types.

We used the SimpleElastix library [145] in Python to implement this co-registration algorithm. Table 4.1 lists the key registration components. For every step, we follow a hierarchical multi-resolution strategy to gradually increase transformation and data complexity, making the co-registration more robust. In terms of transformation complexity, we increase the degrees of freedom by sequentially applying a rigid, affine, and (if applicable) B-spline transformation. To avoid unrealistic deformations, the B-spline deformation is regulated using a bending energy penalty, and is gradually refined across 4 grid levels with increasing density of control points. In terms of data complexity, we apply Gaussian smoothing to the images to create four resolution levels. For every level, 500 co-registration iterations are performed with adaptive stochastic gradient descent (ASGD) as optimiser and linear interpolation. To speed up the process, a random coordinate sampler is used to calculate the metrics. Full co-registration parameter files are found in Appendix A.

Table 4.1. Registration components for the first three steps of the co-registration algorithm.

Component	Step I	Step II	Step III
Moving Image	Micro-CT fresh	Registered micro-CT fresh	Selected slice micro-CT fresh
Fixed Image	Micro-CT fixated	WSI	WSI
Transformation	Rigid>Affine>B-spline	Rigid>Affine	Rigid>Affine>B-spline
Metric	NCC	NMI	NMI
Optimiser	ASGD	ASGD	ASGD
Interpolator	Linear	Linear	Linear

WSI = whole slide image, NCC = normalised correlation coefficient, NMI = normalised mutual information, ASGD = adaptive stochastic gradient descent.

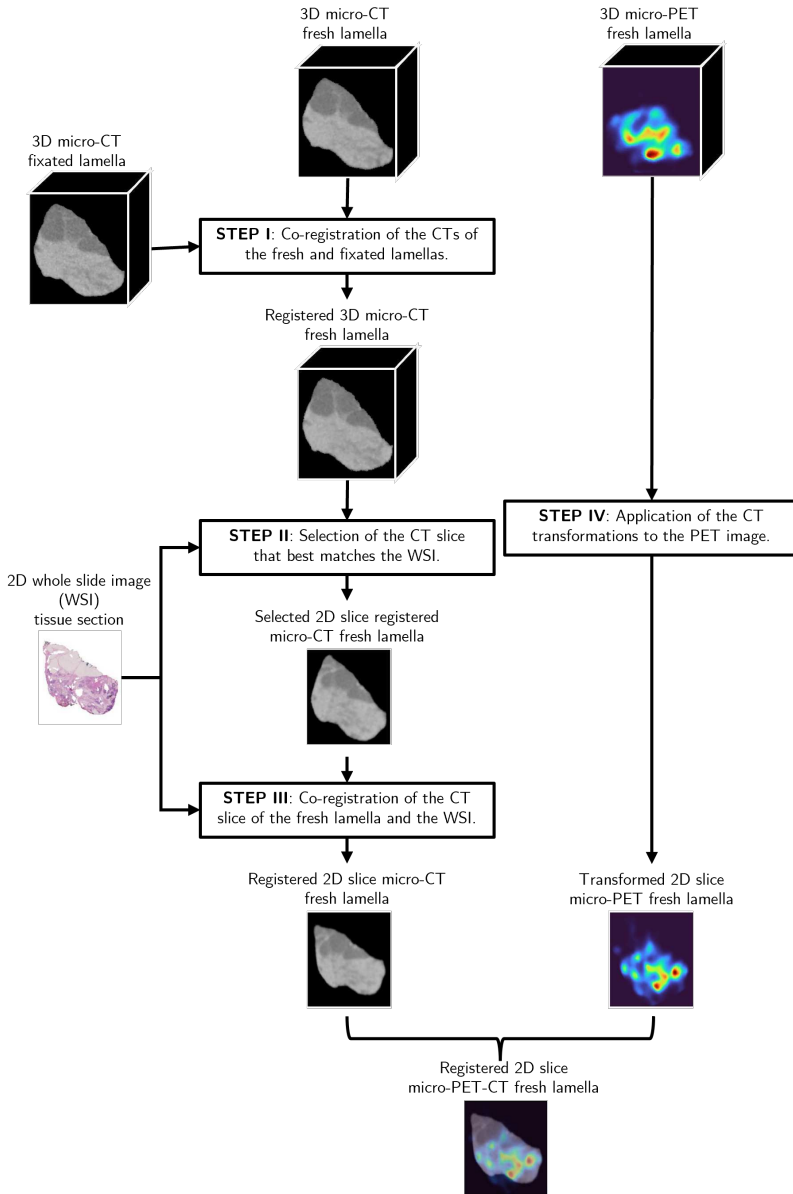


Figure 4.2. Illustration of the four-step co-registration algorithm.

4.2.2. Application and evaluation in breast cancer

4.2.2.1. Study design and subjects

We applied and evaluated the proposed co-registration method in BCa. We acquired images of specimens from 28 patients with NST, ILC, or DCIS BCa undergoing breast-conserving surgery at Ghent University Hospital. Specimens included 19 NST, 6 ILC, and 3 DCIS cases, and we imaged 53, 19, and 10 cassettes of these specimens, respectively. Patients received a preoperative injection with either 0.8 MBq/kg or 4 MBq/kg [^{18}F]FDG. Data collection occurred within 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. All patients provided written informed consent, agreeing to study participation and publication of the results.

4.2.2.2. Image acquisition

All images were captured following the image acquisition scheme proposed in Section 4.2.1.1. After the specimen was cut into lamellas, the pathologist visually assessed which lamellas contained macroscopic tumour tissue. From every specimen, at least one lamella with and one without macroscopic tumour were selected. Large lamellas were further cut to fit the cassettes. Figure 4.3 provides an overview of all images that were collected for one lamella. The following sections explain in detail how these images were acquired.

Micro-PET-CT imaging of lamellas We acquired micro-PET-CT images of the cassettes containing the fresh lamellas. After fixating the tissue in a 10% neutral buffered formalin solution for 12-18 hours, we also acquired micro-CT images of the cassettes with the fixated lamellas. All micro-PET-CT images for this study were either acquired with the β -CUBE[®] and X-CUBE[®] (MOLECUBES, Belgium), or the AURA[®]10 (XEOS, Belgium). While the β -CUBE[®] and X-CUBE[®] are meant for preclinical *in vivo* imaging, and the AURA[®]10 is dedicated to *ex vivo* specimen scanning, both scanners have similar specifications, and offer similar image quality with a submillimetre spatial resolution for PET. Only one micro-PET-CT and one micro-CT acquisition per patient was required, as the bed or container of these scanners could simultaneously fit all cassettes. To obtain a uniform dataset, 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) algorithm, with an energy window of 50% around the 511 keV photopeak and an isotropic voxel size of 400 μm . CT acquisition took 3 min with the X-CUBE[®] and 1 min with the AURA[®]10. CT images

were reconstructed using the image space reconstruction algorithm (ISRA) (X-CUBE[®]) or filtered backprojection (FBP) (AURA[®]10), with an isotropic voxel size of 100 μm . The result was a set of co-registered micro-PET and micro-CT images in DICOM format. The 3D micro-CT images were expressed in Hounsfield units (HUs) and the 3D micro-PET images were converted to lean body mass SUVs using the Janmahasatian formulation [92]. All micro-PET images were attenuation corrected and median filtered (kernel size of 5 $\times$ 5 $\times$ 5 voxels) to remove noise.

Whole slide imaging of H&E stained tissue sections After micro-PET-CT imaging, the cassettes containing the fixated tissue underwent standard histopathological processing to obtain one H&E stained tissue section per cassette. Digital WSIs were captured from these sections using a slide scanner (NanoZoomer[®], Hamamatsu Photonics, Japan), resulting in 2D RGB images in NDPI format with a resolution of 0.456 μm .

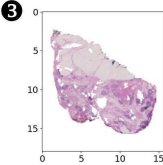
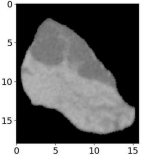
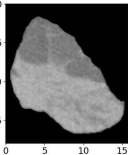
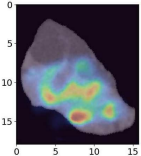
	Fresh breast lamella	Fixated breast lamella	Pathology slide
Optical Photo			
Micro-CT			
Micro-PET-CT			

Figure 4.3. Overview of images collected for one breast lamella. (1) Micro-PET-CT image of the fresh lamella, (2) Micro-CT image of the lamella after formalin fixation, and (3) Digital whole slide image of the haematoxylin and eosin stained thin tissue section. Only one slice is shown for the 3D micro-PET and micro-CT images. The spatial scale of the images is expressed in mm.

4.2.2.3. Mathematical co-registration

Before co-registration, the collected images required preprocessing. We used the `resize` function from the Python library OpenCV [146] to interpolate the micro-PET images and WSIs to the same isotropic voxel size as the micro-CT images (100 μm). A linear (“INTER_LINEAR”) interpolation was applied to upsample the micro-PET images, whereas area-weighted averaging (“INTER_AREA”) was used to downsample the WSIs [146]. In area-weighted averaging, each output pixel is the average of the corresponding input pixels weighted by their fractional area of overlap with the output pixel. Using AMIDE [147], we cropped the micro-PET-CT images to separate the different tissue samples and to remove the cassettes from the micro-CT images. We segmented the tissue in the micro-CT images and WSIs to define binary tissue masks. For the micro-CT images, we applied Otsu’s thresholding and selected the largest connected tissue structure using OpenCV. For the WSIs, we manually segmented the largest connected tissue structure using QuPath [148]. The resulting tissue masks were used to remove the background from the micro-CT images and WSIs, and to crop the images with a margin of 5 voxels around the tissue in all directions. We padded all images that belonged to one lamella to the same dimensions, and we converted the WSIs to greyscale images. Finally, the WSIs were rotated manually to obtain an approximate initial alignment with the micro-PET-CT images. After preprocessing the data, the four-step co-registration algorithm introduced in Section 4.2.1.2 was applied.

4.2.2.4. Co-registration evaluation

The alignment between the transformed micro-PET-CT images of the fresh breast lamellas and the WSIs was evaluated qualitatively through visual inspection and quantitatively using two approaches. For the first approach, we calculated the overlap between corresponding tissues in both images using the DSC, as done by Schiller et al. [123] and de Boer et al. [127]. We calculated the DSC for two types of tissue segmentations: tissue shape, and tissue composition segmentations. For the tissue shape segmentations, we used the binary tissue masks of the micro-CT images and WSIs after co-registration. For the tissue composition segmentations, we distinguished between adipose and non-adipose tissue, as these two tissue types can be separated both in micro-CT images and in WSIs in case of breast tissue. To separate both tissues in micro-CT images, we applied Otsu’s thresholding to all tissue voxels, after clipping the voxel intensities to $[-1000:1000]$ HUs. The WSIs show the greatest contrast between adipose and non-adipose tissue in the green channel. In QuPath, this channel was manually thresholded after applying the binary tissue mask and a Gaussian filter with a sigma of 1 pixel. To calculate the DSC for the composition segmentation, we averaged the DSCs for adipose and non-adipose tissue. The DSC scores based on these tissue shape and composition segmentations are here referred to as $\text{DSC}_{\text{shape}}$ and DSC_{comp} .

As a second evaluation method, we calculated the spatial co-registration error, using an approach similar to Puri et al. [124, 125]. To obtain a mean overall co-registration error between micro-PET and histopathology, we calculated the root sum square (RSS) of the following three errors:

(1) Error between micro-PET and micro-CT of fresh lamellas To estimate the misalignment between consecutively acquired micro-PET and micro-CT images, we averaged the co-registration errors of six different micro-PET-CT scanners, measured using an in-house developed spatial calibration phantom.

(2) Error between micro-CTs of fresh and fixated lamellas To estimate the misalignment between co-registered micro-CTs of the fresh and fixated lamellas (step I), we calculated the distances between pairs of landmarks in both images. As landmarks, we used the microcalcifications that are visible in the micro-CT images. First, we segmented the microcalcifications in the micro-CT images of the fresh lamellas. We thresholded the images at 400 HUs to identify dense structures. The thresholded images were manually corrected to exclude dense spots of histological marking ink. Next, we identified the corresponding microcalcifications in the micro-CT images of the fixated lamellas following the same approach. Per lamella, we calculated the co-registration error as the root mean square (RMS) of all distances between centres of gravity of corresponding microcalcifications.

(3) Error between micro-CT slices of fresh lamellas and WSIs To estimate the misalignment of co-registered selected micro-CT slices of fresh lamellas and WSIs (step III), we could again employ the microcalcifications as they are also visible on the WSIs. We identified the remaining microcalcifications on the selected micro-CT slices, and a breast pathologist with over 15 years of experience segmented the corresponding microcalcification in the WSIs using QuPath. We selected (clusters of) microcalcifications with a minimal diameter of 100 μm to account for the micro-CT resolution. Per lamella, we again calculated the co-registration error as the RMS of all distances between centres of gravity of corresponding microcalcifications. Note that this method could not address possible mismatches between WSIs and selected micro-CT slices (step II).

4.2.2.5. Standardized uptake values

The co-registration method establishes a direct mapping between micro-PET-CT images and the corresponding WSIs, facilitating the analysis of radiotracer uptake in different histological tissue types. An experienced breast pathologist annotated the breast WSIs. Using QuPath, (pre)malignant tissue was delineated, distinguishing between different histological subtypes: NST, DCIS, ILC, and lobular carcinoma in situ (LCIS). Other types of benign and healthy tissue that were annotated are

hyperplasia, inflamed tissue, healthy glands, and skin. The remaining tissue was separated into adipose and connective tissue, using a threshold as described above in Section 4.2.2.4.

We used these annotations together with the co-registered micro-PET images to calculate the mean and maximum SUVs (SUV_{mean} and SUV_{max}) of different breast tissues. In order to do this, we assigned one tissue label to every PET pixel. The WSI annotations were exported with an isotropic pixel size of $3.6 \mu\text{m}$. To interpolate these annotations to the $100 \mu\text{m}$ pixel size of the co-registered micro-PET images, we used a method that accounts for the spatial resolution of micro-PET. The signal of a PET pixel can be seen as a weighted sum of signals originating from different tissues surrounding the PET pixel. The relative contribution of a tissue signal to the PET pixel signal decreases with distance from the PET pixel. These relative contribution weights can be represented by a normalised 2D Gaussian kernel that is centred on the PET pixel. In correspondence with the resolution of the micro-PET scanner, we used a Gaussian kernel with a full width at half maximum (FWHM) of 1 mm and a size ranging from -3σ to $+3\sigma$ in both directions. To every PET pixel, we assigned the label of the tissue type that has the highest total relative weight, and as such contributes the most to the PET pixel signal. In practice, this can be seen as a filtering process. The Gaussian kernel was scanned over the full resolution annotation with a stride of $100 \mu\text{m}$ in both directions. At every kernel position, corresponding to one PET pixel, we selected the tissue type with the highest total relative weight according to the Gaussian kernel. This filtering process is illustrated in Figure 4.4.

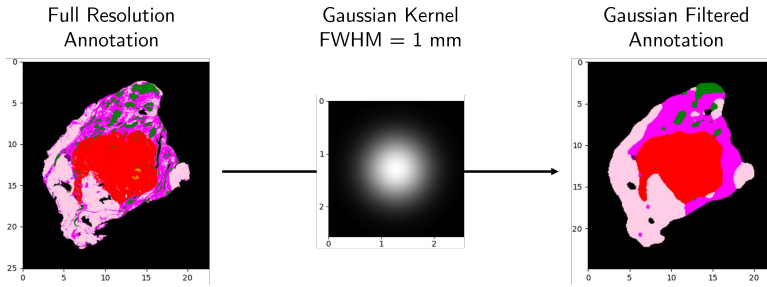


Figure 4.4. Illustration of the Gaussian filtering process of a tissue annotation. The figure illustrates how a full resolution annotation with an isotropic pixel size of $3.6 \mu\text{m}$ is turned into a Gaussian filtered annotation with an isotropic pixel size of $100 \mu\text{m}$. The Gaussian kernel with full width at half maximum (FWHM) of 1 mm is scanned over the original annotation with a stride of $100 \mu\text{m}$, and at every position the label of the tissue type with the highest relative weight is chosen. The spatial scale of the images is expressed in mm. Note that the Gaussian kernel is not displayed in its actual relative size.

4.3. Results

4.3.1. Image acquisition in breast cancer

Micro-PET-CT imaging of the fresh breast lamellas was successful in all cases. The mean time between [^{18}F]FDG injection and micro-PET acquisition was $4\text{h}37\text{min} \pm 53\text{min}$ across all patients. Micro-CT images of the lamellas after formalin fixation were only captured for the first 22 patients. Comparing the 3D binary tissue masks of the micro-CT images before and after fixation for these 22 patients revealed an average tissue volume increase of $0.8 \pm 3.9\%$ relative to the volume of the fresh lamellas across all patients. Thus, some lamellas slightly increased in volume, while others slightly decreased. We did not observe any relation between tissue volume increase during fixation and tissue composition in terms of the volume ratio of adipose and non-adipose tissue in a lamella. We have to keep in mind that these results might be affected by partial volume effects in the micro-CT images.

Encouraged by these findings, an intermediate analysis of the collected data was performed to investigate the value of the micro-CT of the fixated lamella in the co-registration process. It was found that including the co-registration of the micro-CTs of the fresh and fixated lamellas (step I), as opposed to omitting this first step, led to minimal differences in the final result. Appendix B shows a detailed qualitative and quantitative comparison between co-registration results with and without guidance of the micro-CT of the fixated lamellas. As collecting micro-CT images of the fixated lamellas was labour-intensive, it was chosen to omit this acquisition for the final six patients.

On average, H&E stained pathology slides were available three days after the specimen arrived at the pathology lab. WSIs of the pathology slides could be acquired for all lamellas. A clear correlation between the micro-PET-CT images and histopathology was immediately visible, due to the cassette framework that minimised tissue deformation.

4.3.2. Mathematical co-registration in breast cancer

Mathematical co-registration of the micro-CT images of the fresh lamellas with the WSIs was performed fully automatically (after preprocessing). The entire co-registration pipeline (steps I, II, III, and IV) took on average 6 ± 2 min across all lamellas of the first 22 patients when run on an 11th Gen Intel Core i7 CPU. For the last 6 patients the co-registration pipeline (steps II, III, and IV) took on average 8 ± 3 min across all lamellas. The intermediate results of the co-registration process are visualised for one lamella in Figure 4.5. Table 4.2 displays quantitative metrics to show the

improvement in image alignment after the different co-registration steps. The top part of the table (Co-registration of the micro-CTs of the fresh and fixated lamellas) contains the data for all lamellas of the first 22 patients, while the bottom part (Co-registration of the micro-CT slice of the fresh lamella and the WSI) contains the data for all lamellas of all 28 patients.

The found transformations were applied to the micro-PET images of the fresh lamellas. Figure 4.6 shows the resulting alignment of the micro-PET-CT images of the fresh lamellas and the corresponding WSIs for some representative NST, ILC, and DCIS cases. The mathematical co-registration further optimised the alignment between the micro-PET-CT images and WSIs, showing a good correspondence between both imaging modalities in all cases.

To estimate the mean overall spatial co-registration error between micro-PET and histopathology, we calculated the RSS of three co-registration errors. The mean co-registration error between the micro-PET and micro-CT images of the fresh lamellas was 0.35 ± 0.06 mm. In the micro-CT images of the fresh lamellas, we identified 64 microcalcifications across 20 lamellas of 17 patients. For 14 of these patients, accounting for 31 microcalcifications across 15 lamellas, micro-CT images of the fixated lamellas were acquired. For these patients, corresponding microcalcifications were easily identified in the micro-CT images of the fixated lamellas. After co-registration of the micro-CTs of the fresh and fixated lamellas (step I), the mean co-registration error was 0.21 ± 0.26 mm across the 15 lamellas. After selecting micro-CT slices matching the WSIs (step II), 6 microcalcifications across 5 slices from 5 different patients remained. The corresponding microcalcifications were found in the WSIs for 5 out of 6 microcalcifications. The WSI without a corresponding microcalcification showed a void, suggesting the microcalcification was lost during slide preparation. After co-registration of the micro-CT slices and WSIs (step III), the mean co-registration error was 0.49 ± 0.33 mm across the 4 slices with corresponding microcalcifications. The resulting RSS of these three errors was 0.69 ± 0.23 mm, averaged across the 4 lamella slices for which all three errors could be calculated. Figure 4.7 visualises the centres of gravity of corresponding microcalcifications in the micro-CT images and WSIs for one lamella.

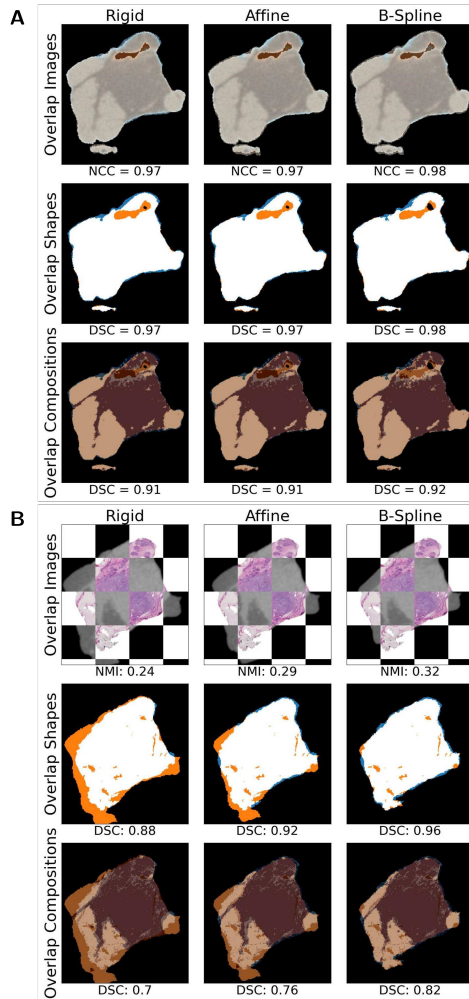


Figure 4.5. Intermediate results of the mathematical co-registration process for one breast lamella. (A) Alignment between the micro-CTs of the fresh and fixated lamella after rigid, affine and B-spline co-registration. Colour fusions show the overlap between both micro-CT images, as well as their shape and composition segmentations. The overlap metrics (normalised correlation coefficient (NCC) and Dice similarity coefficient (DSC)) are shown below the images. Only a central slice of the 3D micro-CT images is shown. (B) Alignment of the selected micro-CT slice of the fresh lamella with the histopathology whole slide image (WSI) after rigid, affine and B-spline co-registration. Chequerboards show the overlap between the micro-CT image and WSI, and colour fusions show the overlap between their shape and composition segmentations. The overlap metrics (normalised mutual information (NMI) and DSC) are shown below the images.

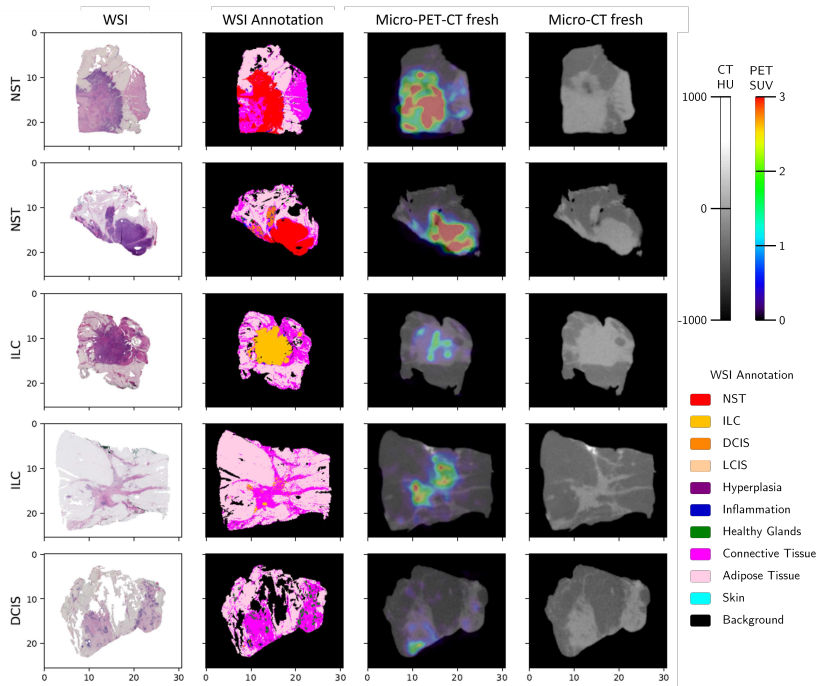


Figure 4.6. Overview of the alignment between the micro-PET-CT images and histopathology whole slide images (WSIs) after mathematical co-registration. Representative lamellas containing different breast tumours are shown: 2 carcinomas of no special type (NST), 2 invasive lobular carcinomas (ILCs), and 1 ductal carcinoma in situ (DCIS). For each lamella the following images are shown: the WSI, the WSI annotation, the co-registered micro-PET-CT image of the fresh lamella, and the micro-CT image without the micro-PET overlay. The original unfiltered WSI annotations are shown. Micro-PET images are displayed with an absolute window of 0-3 standardised uptake values (SUVs) and micro-CT images are shown in Hounsfield units (HUs). The spatial scale of the images is expressed in mm.

Table 4.2. Quantitative metrics expressing the image alignment after the different co-registration steps. The mean $\pm$ standard deviation of all quantitative metrics across the different lamellas are shown after every co-registration step.

	NCC	NMI	DSC _{shape}	DSC _{comp}
Co-registration of the micro-CTs of the fresh and fixated lamella				
Unaligned	0.86 ± 0.06	/	0.87 ± 0.06	0.69 ± 0.11
Rigid	0.92 ± 0.04	/	0.93 ± 0.04	0.78 ± 0.09
Affine	0.94 ± 0.03	/	0.94 ± 0.02	0.80 ± 0.08
B-spline	0.97 ± 0.01	/	0.97 ± 0.01	0.83 ± 0.06
Co-registration of the micro-CT slice of the fresh lamella and the WSI				
Unaligned	/	0.18 ± 0.04	0.79 ± 0.11	0.54 ± 0.15
Rigid	/	0.20 ± 0.04	0.82 ± 0.11	0.58 ± 0.16
Affine	/	0.25 ± 0.04	0.89 ± 0.07	0.63 ± 0.14
B-spline	/	0.27 ± 0.04	0.92 ± 0.06	0.66 ± 0.15

WSI = whole slide image, NCC = normalised correlation coefficient, NMI = normalised mutual information, DSC = Dice similarity coefficient

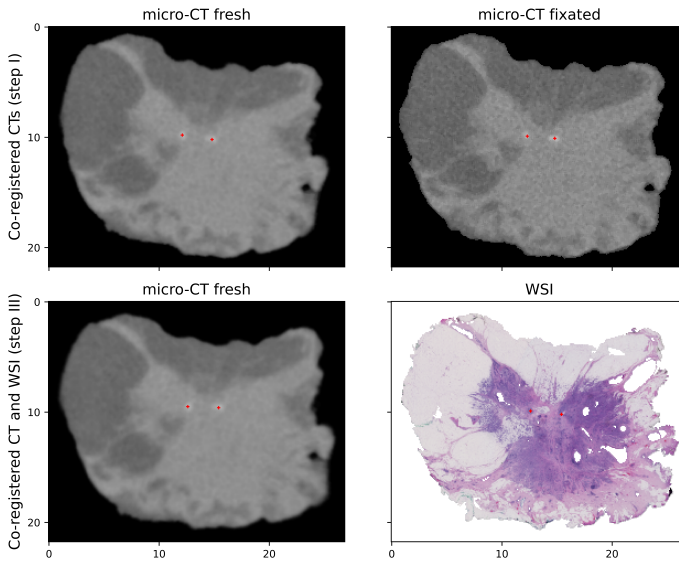


Figure 4.7. Illustration of corresponding microcalcifications in micro-CT images and histopathology whole slide images (WSIs). The top row shows one slice of the aligned micro-CTs of the fresh and fixated lamellas after co-registration step I. The bottom row shows the aligned micro-CT slice and WSI after co-registration step III. Red crosses indicate the centres of gravity of two microcalcifications in all images. The spatial scale of the images is expressed in mm.

4.3.3. Standardised uptake values in breast cancer

After co-registration, we used the histopathology annotations to calculate the SUVs of different breast tissues. Examples of these annotations are shown in Figure 4.6, together with the aligned micro-PET-CT images. Table 4.3 shows the SUV_{mean} and SUV_{max} values per breast tissue type, as defined by the Gaussian filtered annotations. The average and standard deviation of SUV_{mean} and SUV_{max} were calculated across patients. The box plots in Figure 4.8 give a more detailed view on the spread in SUV_{mean} and SUV_{max} across all patients. While LCIS was detected in several lamellas, no LCIS regions remained after Gaussian filtering of the annotations. Hence, no results on the SUVs of LCIS could be reported. To evaluate how well different breast tissue types can be distinguished from healthy tissue, we calculated the uptake of each tissue type relative to that of healthy tissue per patient. Table 4.4 shows the ratios of SUVs of different breast tissue types relative to healthy tissue, here defined as the combination of adipose, connective, and healthy glandular tissue. The averages and standard deviations of the ratios of SUV_{mean} and SUV_{max} were calculated across patients. The box plots in Figure 4.9 give a more detailed view on the spread in SUV_{mean} and SUV_{max} ratios across all patients.

Table 4.3. Standardised uptake values (SUVs) of different breast tissue types as measured in [^{18}F]FDG micro-PET images. The mean $\pm$ standard deviation of SUV_{mean} and SUV_{max} across all patients are shown. The number of patients included in each average is indicated in parentheses.

	SUV_{mean}	SUV_{max}
NST (19)	3.65 $\pm$ 4.32	9.32 $\pm$ 8.40
ILC (6)	0.56 $\pm$ 0.48	1.88 $\pm$ 1.03
DCIS (3)	0.92 $\pm$ 0.4	1.75 $\pm$ 1.75
LCIS (0)	-	-
Hyperplasia (1)	0.04	0.09
Inflammation (9)	1.54 $\pm$ 1.84	2.96 $\pm$ 4.27
Healthy glands (7)	0.21 $\pm$ 0.15	0.74 $\pm$ 0.62
Connective tissue (28)	0.50 $\pm$ 0.58	4.14 $\pm$ 3.26
Adipose tissue (28)	0.11 $\pm$ 0.10	3.63 $\pm$ 4.85
Skin (1)	2.99	6.80

NST = carcinoma of no special type, ILC = invasive lobular carcinoma, DCIS = ductal carcinoma in situ, LCIS = lobular carcinoma in situ

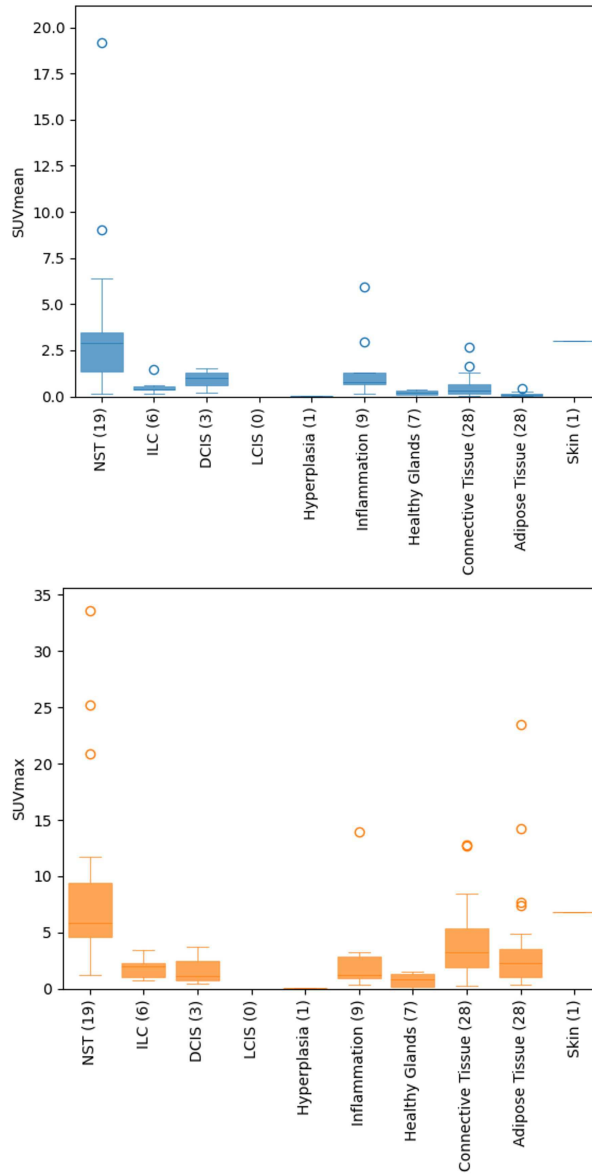


Figure 4.8. Box plots showing the spread in standardised uptake values (SUVs) of different breast tissue types. The top plot shows the spread in SUV_{mean}, while the bottom plot shows the spread in SUV_{max} across all patients. The number of patients included in each box plot is indicated in parentheses.

Table 4.4. Ratios of standardised uptake values (SUVs) of different breast tissue types relative to healthy tissue, as measured in [¹⁸F]FDG micro-PET images. Healthy tissue is defined as combined adipose, connective, and healthy glandular tissue. The mean ± standard deviation of the SUV_{mean} and SUV_{max} ratios across all patients are shown. The number of patients included in each average is indicated in parentheses.

	SUV _{mean}	SUV _{max}
	SUV _{mean} Healthy Tissue	SUV _{max} Healthy Tissue
NST (19)	7.78 ± 5.82	1.87 ± 1.00
ILC (6)	6.18 ± 5.89	1.82 ± 1.13
DCIS (3)	2.18 ± 0.58	0.38 ± 0.26
LCIS (0)	-	-
Hyperplasia (1)	0.09	0.02
Inflammation (9)	3.37 ± 1.41	0.60 ± 0.47
Skin (1)	1.12	0.53

NST = carcinoma of no special type, ILC = invasive lobular carcinoma, DCIS = ductal carcinoma in situ, LCIS = lobular carcinoma in situ

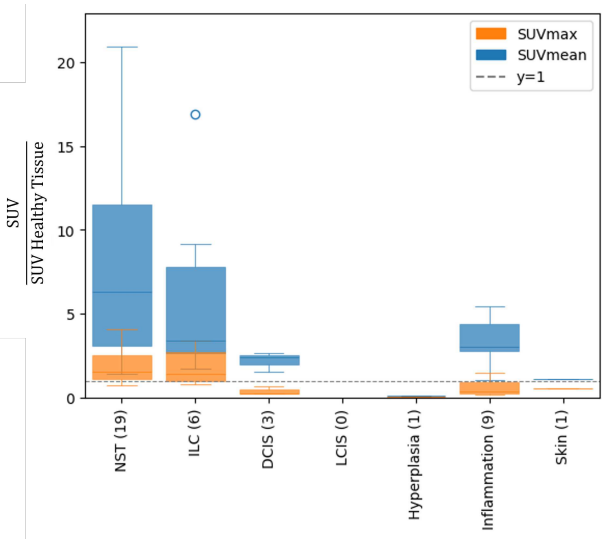


Figure 4.9. Box plots showing the spread across patients of ratios of standardised uptake values (SUVs) for different breast tissue types relative to healthy tissue. Per patient, the ratios of SUV_{mean} and SUV_{max} were calculated per tissue type with respect to healthy tissue (combined adipose, connective, and healthy glandular tissue). The number of patients included in each box plot is indicated in parentheses.

4.4. Discussion

We developed a method for co-registration of submillimetre PET images of *ex vivo* tissue lamellas with the corresponding histopathology. While the idea of imaging sliced tissue specimens as an intermediate co-registration step was already introduced by Puri et al. [125], our method has several advantages over existing co-registration methods that align *in vivo* PET images with histopathology [123, 124, 126–129]. First, the only requirement of our method, in addition to the standard histopathological processing, is that the tissue should be sliced in lamellas while it is still fresh. Second, our mathematical co-registration method is fully automatic as we use intensity-based similarity metrics based on the entire images. Finally, *ex vivo* PET imaging of tissue allows visualising radiotracer uptake at a submillimetre resolution that cannot yet be reached by *in vivo* PET scanners.

We applied the proposed method to resected BCa specimens. The micro-PET-CT images and the corresponding WSIs showed good visual alignment after co-registration. Quantitative metrics (NCC, NMI, and DSC scores) improved with more complex co-registration transformations. This indicates the need to correct for free-form deformations to reach accurate alignment, consistent with the conclusions of de Boer et al. [127]. Schiller et al. [123] reported a maximal DSC of 0.90 ± 0.02 for the overlap between the prostate shapes in *in vivo* CT and histopathology. A maximal DSC of 0.94 ± 0.02 was reported by de Boer et al. [127] for the overlap between the breast tissue shapes in *ex vivo* white light images and histopathology. While not strictly comparable, we found a similar maximal DSC of 0.92 ± 0.06 for the overlap between the breast tissue shapes in *ex vivo* micro-CT slices of the fresh lamellas and WSIs. The DSC_{comp} scores were lower than the DSC_{shape} scores. This is most likely influenced by the fact that CT images do not always show high contrast between adipose and non-adipose tissue, resulting in small inaccuracies in the composition segmentations. Using the microcalcifications as landmarks, we found a mean overall co-registration error between micro-PET and histopathology of 0.69 ± 0.23 mm, which is comparable to the spatial resolution of the used specimen micro-PET scanner. This error is also smaller than the co-registration error of 3.0 ± 0.7 mm reported by Puri et al. [124, 125] for the alignment between *in vivo* PET images and histopathology slices in laryngeal cancer. These results show that the proposed co-registration method is successful, even for soft tissue like breast tissue.

The co-registration method allowed us to collect a unique dataset of [^{18}F]FDG micro-PET-CT images of lumpectomy specimens with ground-truth breast tumour segmentations, required for the development of DL-based segmentation models. Furthermore, this dataset also provided valuable insights into the uptake of [^{18}F]FDG in different types of breast tissue at a high resolution. According to the resulting SUV_{mean} values,

(pre)malignant tissue (including NST, ILC, and DCIS) in general showed higher uptake compared to healthy tissue (including healthy glandular, connective, and adipose tissue). ILC and DCIS showed lower uptake compared to NST, which corresponds to what is reported in literature for *in vivo* [^{18}F]FDG PET [88, 141, 143, 149, 150]. While it is often difficult to detect ILC and DCIS in *in vivo* PET, our results suggest that micro-PET-CT images can visualise ILC and DCIS tumours. Hyperplasia showed low uptake, in the same range as healthy tissue. As expected, we also found notable uptake of [^{18}F]FDG in inflamed tissue and in the skin. It is known that [^{18}F]FDG suffers from low specificity and can show non-specific uptake in inflamed tissue, in benign hyperplasia, in sebaceous glands in the skin, and even in healthy breast glands [88, 141, 149].

In terms of SUV_{max} values, our results follow the same trend compared to values reported by other groups for *in vivo* [^{18}F]FDG PET. Fujii et al. found mean SUV_{max} values of 3.44 ± 3.03 and 2.25 ± 1.99 for NST and ILC respectively [143]. Azuma et al. found a mean SUV_{max} of 2.9 ± 0.5 for the DCIS subtype [150]. Uptake of normal breast tissue is in general found to be homogeneous, with SUV_{max} below 2.5 [149, 151]. The specimen SUV_{max} values we found for BCa of the NST type are clearly higher compared to the values reported for *in vivo* PET. A possible reason for this is that the improved spatial resolution of *ex vivo* PET reduces the partial volume effect [152]. In addition, specimen PET does not suffer from decreased SUVs caused by respiratory motion [152].

We found a wide range of BCa SUVs. In addition to the histological BCa type, there are other factors that influence tumour SUVs. For example, the molecular subtype and grade of the tumour also influence [^{18}F]FDG uptake, as well as the tumour cell density, the microvessel density, and the level of tumour infiltrating immune cells [88, 141, 143, 149, 150, 153–155]. In future work, we could further investigate the relation between breast tumour characteristics and radiotracer uptake.

The proposed co-registration method has many other applications. It bridges the gap between high-resolution molecular imaging and histopathology, stimulating the novel field of transpathology as introduced by Tian et al. [156]. The method provides a unique opportunity to study the distribution of existing and new radiotracers in different types of human tissue at a submillimetre resolution. These insights will help nuclear medicine physicians to correctly interpret clinical PET scans, in particular when the resolution of PET scanners improves. The co-registered micro-PET-CT images and WSIs also allow gaining insight into the tumour microenvironment and the underlying causes of heterogeneity in tumour metabolism. These findings may eventually result in more personalised patient care.

While the co-registration method shows good results, there are some limitations. First, the micro-CT images and the corresponding WSIs can never be matched perfectly due to artefacts in the WSIs. During slicing and colouring of the tissue sections, tissue folds and tears can occur, and small tissue pieces can break off [157]. These artefacts can lead to strong local deformations in the micro-CT images during co-registration. Second, the poor soft tissue contrast of the CT images limits the co-registration accuracy. Third, in the proposed method we make the assumption that the tissue section in the WSI can be matched with a section of the fixated lamella, parallel to the cassette surface. This is an approximation as the tissue might slightly deform when embedding it in paraffin. In addition, cutting may not be performed perfectly parallel to the cassette surface.

Second, the strategy we used to evaluate the spatial co-registration error based on pairs of microcalcifications, also has several shortcomings. First, this evaluation method is not entirely independent as we used intensity-based metrics to perform fully automatic co-registration. Second, not all imaged tissue lamellas contained microcalcifications, so we could only estimate the co-registration error for a subset of the data. Third, large microcalcifications can migrate or even get lost during preparation of the histopathology slides, possibly affecting the co-registration error.

Furthermore, the approach we used to calculate breast tissue SUVs also has some limitations. First, as we are comparing a micro-PET image with an isotropic resolution of around 1 mm with a histopathology WSI with a resolution of 0.456 μm and a thickness of around 5 μm , partial volume effects inevitably influence the results. While we perform 2D Gaussian filtering of the WSI annotations to account for the limited in-plane micro-PET resolution, the PET signal is influenced by the 3D tissue configuration. These partial volume effects will lead to partial volume loss, as well as spillover, which may explain why the resulting SUV_{max} values of healthy glandular, connective, and adipose tissue are rather high. As mentioned by Puri et al. [125], obtaining contiguous histopathology slices to account for the 3D tissue configuration would be more accurate, but introduces unwanted complexity in the co-registration strategy. Second, there are also limitations to the accuracy of the annotations. For maximal accuracy, the annotations should be performed on cellular level, which is not feasible to do manually. Third, the time between radiotracer injection and imaging was quite large, which might have increased the image noise levels. However, in general, visual inspection of the images showed clear PET signals.

Finally, while the proposed co-registration method is considerably less labour-intensive and time-consuming compared to existing methods [123, 124], intermediate image acquisition of fresh and fixated tissue lamellas still involves some manual steps. It was already shown that the proposed workflow can be simplified by omitting the acquisition of micro-CT images of fixated lamellas with minor impact on the final result. In future work,

we aim to develop a method to co-register micro-PET-CT images of full specimens with the corresponding histopathology WSIs, eliminating the need to acquire micro-PET-CT images of tissue lamellas. In addition, such method would be more widely applicable, as it no longer requires fresh cutting of the specimens. A possible approach for such a method would be to combine the methods proposed by Rusu et al. [158, 159] to co-register 3D medical images with a stack of equally spaced histopathology WSIs with our slice matching technique for improved accuracy in high-resolution images.

4.5. Conclusions

We developed a novel method to accurately and automatically co-register high-resolution PET-CT images with histopathology in resected tissue. Our method consists of an image acquisition scheme that minimises and tracks tissue deformation, followed by a mathematical co-registration algorithm to further optimise image alignment. Our method is less labour-intensive and time-consuming compared to previously proposed protocols, making it feasible to study larger patient groups.

We applied our method to BCa specimens and showed that, even for highly deformable breast tissue, co-registration can be successfully performed with an accuracy comparable to the spatial resolution of micro-PET. We collected a unique dataset of [^{18}F]FDG micro-PET-CT images of breast tissue that are co-registered with the corresponding histopathology. This dataset with ground-truth breast tumour segmentations is crucial for the development and validation of DL-based segmentation models. Furthermore, this dataset allowed a detailed study of SUVs of different types of breast tissue. We found that (pre)malignant breast tissue in general shows higher uptake compared to healthy tissue. This not only includes the NST tumours, but also the ILC and DCIS subtypes that are in general more difficult to detect in *in vivo* PET. As expected, high non-specific uptake was also detected in inflamed tissue and in the skin.

We believe that the proposed technique can also be applied to other types of tumours and radiotracers. The proposed method can provide a better understanding of the *in vivo* distribution of (new) radiotracers in different tissues on a submillimetre scale, and may clarify the underlying mechanisms of radiotracer uptake. These insights will eventually improve clinical practice and patient care.

5 | Breast tumour segmentation in micro-PET-CT images of lumpectomy specimens

Abstract

Introduction: Complete tumour removal is vital in curative breast cancer (BCa) surgery to prevent recurrence. Recently, imaging of lumpectomy specimens with high-resolution [^{18}F]fluorodeoxyglucose ([^{18}F]FDG) positron emission tomography (PET), in combination with computed tomography (CT), has shown promise for intraoperative margin assessment (IMA). To aid physicians in interpreting micro-PET-CT images, we aim to develop a decision-support system.

Methods: We trained and evaluated a 2D Residual U-Net (ResU-Net) to delineate (pre)malignant tissue in micro-PET-CT lumpectomy images. For this, we used the dataset of 2D micro-PET-CT images of BCa lamellas with corresponding true histopathology-defined tumour segmentations collected in Chapter 4. In total, the dataset consisted of 82 2D images from 28 patients with invasive carcinoma of no special type (NST), invasive lobular carcinoma (ILC), or ductal carcinoma in situ (DCIS). We compared the performance of the ResU-Net with conventional PET intensity thresholding.

Results: Group five-fold cross-validation yielded a Dice similarity coefficient (DSC) of 0.73 ± 0.20 for segmentation of NST tumours with the ResU-Net. The area under the curve (AUC) of the precision-recall (PR) curve was 0.83 ± 0.14 for the NST tumours, and the ResU-Net outperformed conventional intensity thresholding in terms of the PR-AUC value. Although similar trends were observed for ILC and DCIS, due to the small datasets, we could not show statistically significant improvements compared to intensity thresholding for these BCa subtypes. In addition, we showed that combining micro-PET and micro-CT information is essential for accurate tumour delineation, with dual-modality input models outperforming single-modality input models.

Conclusions: While additional and more diverse datasets are required to further improve and evaluate the segmentation models, overall, the results highlight the ResU-Net's potential as a first step towards a decision-support system to enhance micro-PET-CT-based IMA in BCa.

This chapter contains information previously published in:

Luna Maris, Menekse Göker, Kathia De Man, Bliede Van den Broeck, Sofie Van Hoecke, Koen Van de Vijver, Christian Vanhove, and Vincent Keereman. Supporting intraoperative margin assessment using deep learning for automatic tumour segmentation in breast lumpectomy micro-PET-CT. *npj Breast Cancer, Nature Publishing Group*, 11(1):88, 2025.

5.1. Introduction

Early-stage breast cancer (BCa) is often treated with breast-conserving surgery (BCS), involving local tumour excision [5, 7]. As explained in Section 2.3 of Chapter 2, successful BCS or lumpectomy requires complete removal of the primary tumour with a margin of surrounding healthy tissue [7]. The gold standard to assess the margin status is sectional histopathology [20]. This process involves inking the surface of the resected specimen, and according to the 2014 Society of Surgical Oncology and American Society for Radiation Oncology (SSO-ASTRO) consensus, a positive margin for invasive BCa is defined as “ink on tumour” (i.e., the tumour reaches the specimen surface) [18]. However, this procedure typically requires at least 16 hours, and often even 3-5 days [20]. Following initial excision, positive margins are found in around 20% of patients with invasive BCa, with reported rates of 7%-36% [7, 21, 22]. 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 [15].

To avoid re-surgery, intraoperative margin assessment (IMA) is crucial, allowing immediate removal of additional tissue when margins are suspected to be positive. As previously discussed in Section 2.5 of Chapter 2, specimen imaging with high-resolution (micro-) positron emission tomography (PET), in combination with computed tomography (CT), recently emerged as a promising technique for IMA during BCS [30–32]. For this technique, the radiotracer [^{18}F]fluorodeoxyglucose ([^{18}F]FDG) is preoperatively administered to the patient. Afterwards, during BCS, the resected lumpectomy specimen is visualised in 3D using a dedicated micro-PET-CT specimen scanner, allowing fast margin assessment.

As discussed in Section 2.5.4 of Chapter 2, the BIMAP and BrIMA studies previously investigated how well physicians can interpret micro-PET-CT images of lumpectomy specimens in terms of positive margin detection [30, 32, 33]. In a combined group of 41 lumpectomy specimens of different BCa subtypes, the BIMAP study showed an average sensitivity and equal specificity of 81% across seven physicians postoperatively interpreting the micro-PET-CT images [32]. In the BrIMA study, a sensitivity and specificity of 77% and 67%, respectively, were found for a group of 152 lumpectomy specimens of various BCa subtypes intraoperatively interpreted by the surgeons [33]. Göker et al. [34] reported an average interpretation time of 1 min per 3D micro-PET-CT image and found substantial inter-reader variability for the interpretation of a subset of images from the BIMAP study. In contrast, De Crem et al. [32] found minimal performance differences between physicians with various experience levels after proper interpretation training. In an inter-reader reproducibility study applied to the BrIMA data, high reproducibility was shown when combining interpretations across four surgeons, while variability was seen in single-reader assessments [33].

Although study results look promising, 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 improve sensitivity and specificity. The need for automatic (artificial intelligence-based) algorithms that can enhance micro-PET-CT image interpretation, was previously also expressed by Lambert et al. [31] and Moraitis et al. [35]. In this chapter, we present a deep learning (DL) model that automatically segments (pre)malignant breast tissue in micro-PET-CT lumpectomy images. Such a DL model can form the basis of a clinical decision-support tool.

5.2. Methods

5.2.1. Dataset

To train and test the DL-based breast tumour segmentation models, we collected 2D micro-PET-CT slices of lumpectomy lamellas together with their true histopathological semantic tumour annotations. This dataset was used for supervised training and testing of the DL models. In the previous chapter, Chapter 4, it was discussed in detail how this dataset was collected. We here summarise the most important steps.

5.2.1.1. Study design and subjects

As discussed in Section 4.2.2.1 of Chapter 4, we collected micro-PET-CT images of lumpectomy lamellas from 28 patients with early-stage BCa undergoing BCS at the Ghent University Hospital. Patients were preoperatively injected with 0.8 MBq/kg or 4 MBq/kg [^{18}F]FDG. Imaged specimens included 19 invasive carcinomas of no special type (NST), 6 invasive lobular carcinomas (ILCs), and 3 cases of ductal carcinoma in situ (DCIS). 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 data were collected at the Ghent University Hospital in two clinical trials (ClinicalTrials.gov identifiers: NCT04343079 and NCT04999917).

5.2.1.2. Micro-PET-CT images of breast lamellas

We collected micro-PET-CT images of lumpectomy lamellas. Immediately after resection, the fresh tumour specimens were cut into approximately 2 mm thick lamellas, and two to three lamellas were imaged per patient. All micro-PET-CT images for this study were either acquired with the

β -CUBE[®] and X-CUBE[®] (MOLECUBES, Belgium), or the AURA[®]10 (XEOS, Belgium). To obtain a uniform dataset, all images were reconstructed to emulate a PET acquisition time of 10 min and an injected dose of 0.8 MBq/kg. Image acquisition and reconstruction details are discussed in Section 4.2.2.2 of Chapter 4. The results were 3D co-registered micro-PET and micro-CT images in DICOM format, with an isotropic voxel size of 400 μ m and 100 μ m, respectively. The micro-CT images were expressed in Hounsfield units (HUs) and the micro-PET images were converted to lean body mass standardised uptake values (SUVs) using the Janmahasatian formulation [92]. All micro-PET images were corrected for attenuation and median filtered with a 2 mm isotropic kernel to remove noise.

For further processing, DICOM images were converted to NUMPY arrays. The micro-PET images were linearly interpolated to the grid of the micro-CT images. Micro-CT images were thresholded using Otsu's method to define binary tissue masks [160]. Only the largest connected tissue structures were selected, and the obtained tissue masks were used to remove the background from the micro-CT images and crop the micro-PET-CT images to only include the tissue.

5.2.1.3. Breast tumour segmentation labels

After micro-PET-CT imaging, each lamella was further processed following clinical routine to obtain one haematoxylin and eosin (H&E) stained tissue section, of which a digital WSI was captured (NanoZoomer[®], Hamamatsu Photonics, Japan). The collected micro-PET-CT images and the corresponding WSIs were co-registered using the four-step co-registration algorithm explained in Section 4.2.1.2 of Chapter 4. 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 [148]. As such, we obtained binary tumour annotations, distinguishing between (pre)malignant and benign breast tissue. Lobular carcinoma in situ (LCIS) was included in the benign category, as it does not affect the histopathological margin status. 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.

5.2.2. Deep learning model

5.2.2.1. Architecture: ResU-Net

A commonly used DL model architecture for semantic segmentation is U-Net and its variants [111]. 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 [161]. We used the MONAI [162] implementation of the ResU-Net, as introduced by Kerfoot et al. [113]. 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 [113, 161]. Figure 5.1 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 pixel-wise tumour probabilities ($\in [0,1]$), as output. All code for this work was implemented using PyTorch [163].

5.2.2.2. Data splits, preprocessing, and augmentation

The dataset with tumour segmentation labels 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 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 ranges 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. In addition, the test images were padded with background pixels up to the smallest size suited to pass through the network.

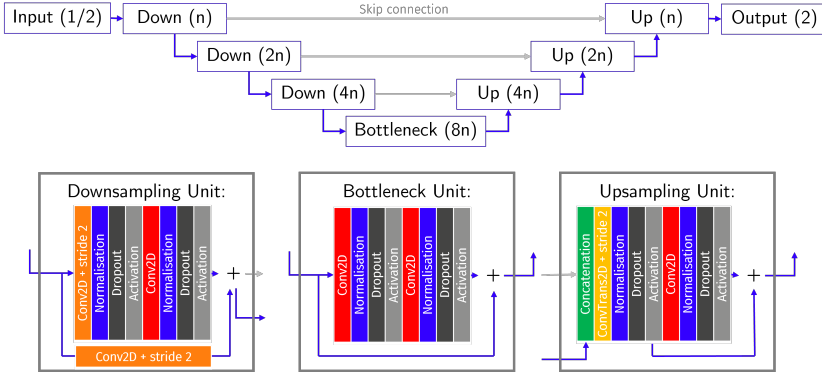


Figure 5.1. Illustration of the Residual U-Net (ResU-Net) architecture used in this work. The composition of the residual downsampling, bottleneck, and upsampling units is shown. The number of layers shown in this figure is only indicative, as a suited number of layers is defined using a grid search. The number of channels (n) is doubled for every downsampling unit and halved for every upsampling unit. We used the MONAI [162] implementation of the architecture introduced by Kerfoot et al. [113].

5.2.2.3. Hyperparameter optimisation

The ResU-Net was initialised using the standard Kaiming initialisation [100]. Parametric ReLU (PReLU) activations, introduced by He et al. [100], were implemented throughout the network, except for the last layer, where a sigmoid activation was used. We used the Dice loss with background exclusion and batch-wise summation of intersection and union, as implemented in MONAI [162]. The stochastic gradient descent (SGD) optimisation algorithm with Nesterov momentum was used [102], 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 Dice similarity coefficient (DSC). To tune other model hyperparameters, we performed a grid search using the Optuna framework [164]. 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 at the centre, 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 results were five hyperparameter combinations, one per test fold.

5.2.2.4. Training

Using the hyperparameters as defined through the 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.

5.2.3. Intensity thresholding

To gain more insight into the value of using a DL model in this segmentation 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 [165–169]. For every train set, we varied the threshold between 0% and 100% of the maximum SUV per patient with a step size of 0.1%, 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 [160]. The obtained high-density regions were then used to mask the micro-PET signal, before defining the most suited 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.

5.2.4. Segmentation performance evaluation

To evaluate the models' ability to correctly segment (pre)malignant breast tissue, we used the trained ResU-Nets and intensity thresholds to predict tumour segmentations for all test cases 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. In case of the ResU-Nets,

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 area under the curve of the precision-recall curve (PR-AUC). The DSC ($\in [0,1]$) scores the overlap between the true and predicted tumour segmentations. It is a well-suited metric to evaluate segmentation performance in medical images with extensive class imbalance [114]. In addition, as we eventually 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 95th percentile of the Hausdorff distance (95HD), as well as the contour Dice with a 1 mm tolerance (1mmCD). For the 1mmCD we implemented a 2D variant of the surface Dice as implemented in MONAI [162]. 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 statistically 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 ($m = 3 \cdot 4 = 12$).

5.3. Results

5.3.1. Dataset

For supervised training and testing of the DL models for segmentation, we obtained 2D micro-PET-CT images of breast lamellas, co-registered with the corresponding histopathology WSIs. The (pre)malignant tissue, here including NST, ILC, and DCIS tumours, was annotated in the WSIs to obtain binary tumour segmentations. Figure 5.2 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.



Figure 5.2. Example of a micro-PET-CT slice with the corresponding true tumour segmentation, as annotated on the histopathological 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 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.

5.3.2. Trained models

Table 5.1 provides an overview of all models trained in this work. 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 joint ILC and DCIS data. 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 Table C.1 of Appendix C. 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. Figures C.1 and C.2 of Appendix C show the evolution of the train and validation loss across epochs for model training on the NST and combined ILC and DCIS cases, respectively.

For the conventional intensity thresholding approach, we used the same stratified group five-fold train-test split to find the most optimal intensity thresholds. 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.

Table 5.1. Overview of all 150 Residual U-Nets (ResU-Nets) and 30 intensity thresholds (Intens thresh) trained in this work, with their respective input images, and the breast cancer (BCa) type they were trained to segment.

Input	BCa 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

NST = carcinoma of no special type, ILC = invasive lobular carcinoma, DCIS = ductal carcinoma in situ

5.3.3. Segmentation performance

To evaluate the models' performance, we used the trained ResU-Nets and intensity thresholds to segment the (pre)malignant tissue in the micro-PET-CT images of all test cases with true segmentation labels. Predicting the tumour segmentation for one 2D micro-PET-CT slice, using the appropriate ensemble of five trained ResU-Nets of the corresponding fold, took less than one second. Figure 5.3 visualises the predicted tumour segmentations for a representative NST test case. By comparing the predicted and true tumour segmentations, we could calculate the DSC, PR-AUC, 95HD, and 1mmCD. Table 5.2 shows the mean scores of the evaluation metrics across the test patients with NST tumours. The box plots in Figure 5.4 show the distribution of the metric values in more

detail for the NST tumours. The tumour segmentation performance of the ResU-Nets and intensity thresholding models was compared using the Wilcoxon signed-rank test. Table 5.2 shows the p-values resulting from this analysis for the NST tumours. Furthermore, 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. The p-values resulting from this analysis for the NST tumours are shown in Table 5.3.

Full test results for ILC and DCIS tumours can be found in Appendix D. 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 Figures D.1 and D.2 of Appendix D, respectively. The mean metric values across the test patients with ILC and DCIS tumours can also be found in Table D.1 of Appendix D. The box plots in Figures D.3 and D.4 of Appendix D show the distribution of the metric values in more detail for the ILC and DCIS tumours, respectively. In addition, Table D.1 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. Furthermore, Table D.2 of Appendix D shows the p-values resulting from the Wilcoxon signed-rank test 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.

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 inputs generally resulted in significantly improved performance in terms of DSC, PR-AUC, and 1mmCD, 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, a significant difference was 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.

Table 5.2. Metric values to evaluate the agreement between true and predicted tumour segmentations for test patients with invasive carcinoma of no special type (NST). The mean $\pm$ standard deviation of the DSC, PR-AUC, 95HD, and 1mmCD across all test patients is shown per model input, 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.

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

DSC = Dice similarity coefficient, PR-AUC = area under the curve of the precision-recall curve, 95HD = 95th percentile of the Hausdorff distance (in mm), 1mmCD = contour Dice with a 1 mm tolerance

Table 5.3. P-values (p) resulting from the Wilcoxon signed-rank test, comparing the performance of the ResU-Nets and intensity thresholding (Intens thresh) models across different model inputs for segmentation of invasive carcinoma of no special type (NST). P-values are reported for each metric and each comparison of model inputs, stratified per segmentation strategy. Significant performance differences following Bonferroni-Holm correction ($\alpha = 0.05$, $m = 12$) are indicated in bold.

BCa type:		NST (N=19)		
Metric	Model	p micro-PET vs. micro-CT	p micro-PET-CT vs. micro-CT	p micro-PET-CT vs. micro-PET
DSC	Intens thresh	0.0323	0.0108	0.0002
PR-AUC	Intens thresh	0.0095	0.0039	0.0002
95HD	Intens thresh	0.0004	0.0002	0.0323
1mmCD	Intens thresh	0.0008	0.0001	0.2101
DSC	ResU-Net	0.0323	0.0028	0.0204
PR-AUC	ResU-Net	0.0446	0.0039	0.0004
95HD	ResU-Net	0.3124	0.0230	0.1956
1mmCD	ResU-Net	0.0401	0.0010	0.0003

DSC = Dice similarity coefficient, PR-AUC = area under the curve of the precision-recall curve, 95HD = 95th percentile of the Hausdorff distance (in mm), 1mmCD = contour Dice with a 1 mm tolerance

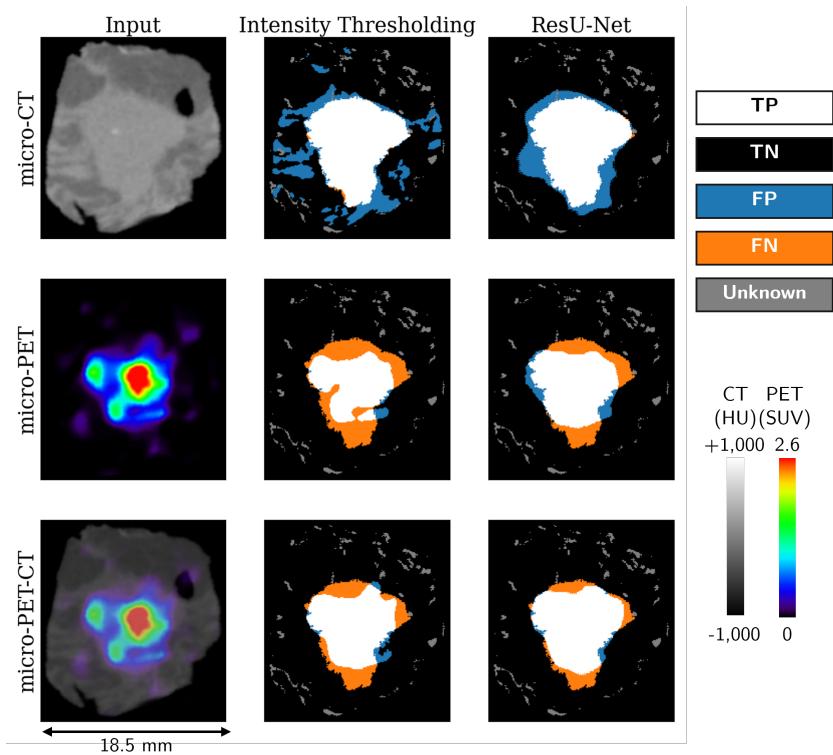


Figure 5.3. Example of the tumour segmentations predicted with intensity thresholding and Residual U-Net (ResU-Net) for an invasive carcinoma of no special type. Predictions are shown for three different inputs: micro-CT, micro-PET, and micro-PET-CT. The micro-PET and micro-CT images are expressed in 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.

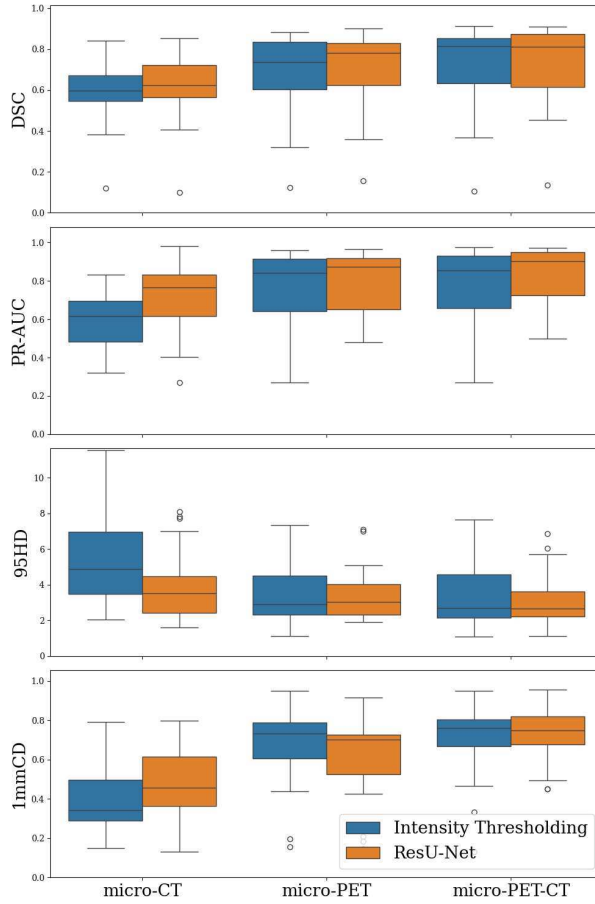


Figure 5.4. Box plots showing metric distributions assessing agreement between true and predicted tumour segmentations for test patients with invasive carcinoma of no special type. The following metrics are shown: Dice similarity coefficient (DSC), area under the curve of the precision-recall curve (PR-AUC), 95th percentile of the Hausdorff distance (95HD) in mm, and 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)).

5.4. Discussion

To our knowledge, we were the first to apply the well-known ResU-Net DL model to perform automatic tumour segmentation in micro-PET-CT images of lumpectomy specimens. We used a dataset 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 0.73 ± 0.20 , and a PR-AUC of 0.83 ± 0.14 for the segmentation of NST breast tumours. 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. While similar trends were observed when studying the ILC and DCIS cases, the limited datasets for these cases did not allow us to show any statistically significant differences in tumour segmentation performance between the ResU-Nets and intensity thresholding models.

Other groups have used DL to automatically segment breast tumours in other types of intraoperative lumpectomy images. Veluponnar et al. [170] and Yeung et al. [171] developed DL models to segment breast tumours in ultrasound (US) images of lumpectomy specimens, while Shia et al. [172] worked on tomosynthesis images, and Kho et al. [173] on hyperspectral images of lumpectomy specimens. Veluponnar et al. [170] reported a median DSC of 0.88 on a set of 109 US images of 86 patients (78 NST and 8 ILC cases), and Yeung et al. [171] reported a mean DSC of 0.96 ± 0.05 on a test set of around 732 US images of three patients (BCa type unclear). Shia et al. [172] reported a mean DSC of 0.86 on a test set of five breast tomosynthesis images of five patients (BCa type unclear). Kho et al. [173] did not report a DSC, but they reached a pixel-level sensitivity of 0.76 ± 0.24 and precision of 0.93 ± 0.09 for a test set of 26 hyperspectral images of 13 tissue slices of 13 patients (10 NST or ILC cases and 3 DCIS cases). Although we obtained a lower DSC of 0.73 ± 0.20 for the NST tumours, the scores are not directly comparable. Except for Kho et al. [173], all groups used manual expert annotations rather than histopathology as gold standard for their segmentation algorithms [170–173]. 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 [135, 168, 169].

This study contains several limitations related to the data. First, there might be slight inaccuracies in the segmentation labels. There might

be imperfections in the tumour annotations created by the pathologist. Amgad et al. [174] reported a DSC of 0.87 when comparing annotations of triple negative BCa made by two senior pathologists. In addition, while in the previous chapter 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 [175]. Second, the datasets used in this work are rather small. Especially for ILC and DCIS tumours, more data is required to enhance and validate ResU-Net performance. While it was found to be challenging to accurately segment ILC and DCIS tumours in micro-PET-CT images of lumpectomy specimens, due to the limited datasets, no meaningful conclusions on model performance could be drawn for these subtypes. 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. Third, we only have 2D micro-PET-CT images with true tumour segmentations available, while the eventual 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.

Before the proposed DL models could be used in clinical practice, additional steps are required. Future research could focus on collecting more 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 evaluated. 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. Such 3D images with segmentation labels could be exploited to train and evaluate a 3D ResU-Net, possibly enhancing model performance. Furthermore, as the ultimate objective is to enable intraoperative margin assessment, it should be investigated how reliable margin status predictions can be derived from the tumour segmentations predicted by the models. In the next chapter, we evaluate the margin assessment performance of the ResU-Nets and the intensity thresholding models. In addition, to further enhance clinical applicability, it should be investigated how trustworthiness of the models can be improved. In Chapter 7, we will investigate how the predicted tumour probability maps, which express spatially-resolved prediction uncertainty, can be calibrated so that the probabilities reflect the true likelihood of malignancy. Finally, we might explore how the DL models can support physicians to improve performance and efficiency for IMA, and we might consider faster, lightweight models for intraoperative use.

5.5. Conclusions

We trained and evaluated a ResU-Net to segment (pre)malignant breast tissue in micro-PET-CT images of lumpectomy specimens. By training and evaluating the model on 2D images with ground-truth histopathology segmentations, we showed that combining micro-PET and micro-CT information is essential for accurate tumour delineation, with dual-modality input models outperforming single-modality input models. For the segmentation of invasive carcinoma of the NST type, the ResU-Net outperformed conventional intensity thresholding in terms of the PR-AUC score. Although similar trends were observed for ILC and DCIS, due to the small datasets, we could not show statistically significant improvements compared to intensity thresholding for these BCa types. Additional work is required before the ResU-Net can be adopted in clinical practice, including the collection of larger and more diverse datasets to further improve and evaluate the segmentation models, especially for ILC and DCIS tumours. In addition, it should be investigated how reliable margin status predictions can be derived from the predicted tumour segmentations, and how model trustworthiness and efficiency can be improved to support intraoperative clinical decision-making. Overall, the results highlight the ResU-Net's potential as a first step towards a decision-support system to enhance IMA in BCa based on micro-PET-CT, while underscoring the need for further (clinical) validation.

6 | Margin status prediction based on micro-PET-CT images of lumpectomy specimens

Abstract

Introduction: Imaging of lumpectomy specimens with high-resolution positron emission tomography (PET), in combination with computed tomography (CT), has shown promise for intraoperative margin assessment (IMA) during breast-conserving surgery (BCS). Our objective is to develop a decision-support system to aid physicians in interpreting these micro-PET-CT images of lumpectomy specimens. In Chapter 5, a 2D Residual U-Net (ResU-Net) was developed to automatically segment (pre)malignant tissue in micro-PET-CT images of lumpectomy specimens. In this chapter, we validate the ResU-Net's ability to accurately predict the margin status of lumpectomy specimens.

Methods: To evaluate margin assessment performance, we applied the ResU-Net to the datasets collected in the BIMAP and BrIMA studies, which contain micro-PET-CT images of lumpectomy specimens together with the histopathological ground-truth margin status. The BIMAP and BrIMA studies collected images of 41 and 152 lumpectomy specimens, respectively, from patients with invasive carcinoma of no special type (NST), invasive lobular carcinoma (ILC), ductal carcinoma in situ (DCIS), or neoadjuvant-treated (NAT) breast cancer (BCa). The ensemble of ResU-Nets trained in Chapter 5 was used to segment (pre)malignant tissue in all images, after which the margin status was predicted based on whether the segmented tumour touched the specimen border. The performance of the ResU-Nets was compared with the performance of conventional PET intensity thresholding models, as well as with the intraoperative and postoperative margin assessments made by physicians interpreting the same micro-PET-CT images.

Results: When comparing the predicted and true histopathological margin status in all NST cases, the ResU-Nets reached an F1 score of 84% and 41% for the BIMAP and BrIMA datasets, respectively. For all BCa subtypes combined, the ResU-Nets reached an F1 score of 76% and 50% for the BIMAP and BrIMA datasets, respectively. Both for the NST tumours and for the combined group of all BCa subtypes in the BIMAP and BrIMA datasets, the ResU-Nets showed margin assessment performance statistically equivalent to intraoperative and postoperative physicians' assessments, while consistently outperforming intensity thresholding. Although statistical significance could not always be shown for smaller ILC,

DCIS, and NAT subsets, the ResU-Nets generally showed higher specificity, precision, and F1 scores compared to intensity thresholding models.

Conclusions: The fully automatic ResU-Net showed a performance comparable to physicians when predicting the margin status of lumpectomy specimens based on micro-PET-CT images. This highlights its potential as a clinical support tool, possibly improving consistency and reducing decision-making time when interpreting micro-PET-CT images. While more extensive validation of the ResU-Net (in a clinical context) is required, this deep learning model, capable of predicting both tumour segmentation and margin status, represents an important step towards a decision-support system that enhances micro-PET-CT-based IMA in BCa, facilitating its clinical adoption.

This chapter contains information previously published in:

Luna Maris, Menekse Göker, Kathia De Man, Bliede Van den Broeck, Sofie Van Hoecke, Koen Van de Vijver, Christian Vanhove, and Vincent Keereman. Supporting intraoperative margin assessment using deep learning for automatic tumour segmentation in breast lumpectomy micro-PET-CT. *npj Breast Cancer*, Nature Publishing Group, 11(1):88, 2025.

6.1. Introduction

Imaging of lumpectomy specimens with high-resolution positron emission tomography (PET), in combination with computed tomography (CT), has recently shown promise for intraoperative margin assessment (IMA) during breast-conserving surgery (BCS) [30, 32]. Although study results for this technique look promising, we hypothesize 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 margin assessment performance [32, 33]. In this chapter, we continue our work on the Residual U-Nets (ResU-Nets) for automatic segmentation of (pre)malignant breast tissue in micro-PET-CT images, introduced in Chapter 5. Building further on these deep learning (DL) models, we here evaluate whether the correct surgical margin status can be derived from the tumour segmentations predicted by these models. As the eventual goal of the decision-support system is to help physicians in the critical task of performing IMA in BCS, validating the margin assessment performance of the proposed DL models is of high clinical relevance.

To this end, we apply the trained ResU-Nets to the 3D micro-PET-CT images of lumpectomy specimens collected in two clinical studies: BIMAP and BrIMA. The BIMAP and BrIMA studies, previously described in Section 2.5.4 of Chapter 2, both include the histopathological ground-truth margin status for all lumpectomy specimens [32, 33]. The predicted margin status derived from the automated tumour segmentation of the ResU-Nets is compared with histopathology, as well as with the margin assessments performed by physicians interpreting the same images intraoperatively and postoperatively in the BIMAP and BrIMA studies. Evaluating how the DL models are performing in comparison with human observers is essential to determine the clinical applicability of these models as decision-support tools.

6.2. Methods

6.2.1. Datasets

We validated the DL models in terms of margin assessment accuracy on the BIMAP [32] and BrIMA [33] datasets, previously discussed in Section 2.5.4 of Chapter 2. Both datasets contain 3D micro-PET-CT images of entire lumpectomy specimens, together with the true histopathological margin status (positive or negative). Table 2.2 of Chapter 2 provides an overview of patients included in the BIMAP and BrIMA studies, along with their breast cancer (BCa) subtypes and ground-truth margin status. We here summarize the most important steps in the collection of both datasets.

6.2.1.1. Study design and subjects: BIMAP and BrIMA

All data to validate the DL models in terms of margin assessment accuracy were collected in the BIMAP (ClinicalTrials.gov identifier: NCT04343079) and BrIMA (ClinicalTrials.gov identifier: NCT04999917) studies. While the BIMAP study was conducted at Ghent University Hospital in Belgium, the BrIMA study was conducted at six European BCa centres in Belgium (n=3), Germany (n=2), and Italy (n=1). Both studies were approved by all local ethics committees and conducted in accordance with the Declaration of Helsinki. All patients provided written informed consent for study participation and publication of the results.

Micro-PET-CT images and histopathological data from specimens of patients with early-stage BCa undergoing BCS were collected. Patients with invasive carcinoma of no special type (NST), invasive lobular carcinoma (ILC), or ductal carcinoma in situ (DCIS), as well as neoadjuvant-treated (NAT) patients were enrolled. Both studies included patients with NST BCa who were pretreated with neoadjuvant chemotherapy (NAC), and the BrIMA study also included cases of NST pretreated with aromatase inhibitors (NST-AI) and ILC pretreated with aromatase inhibitors (ILC-AI). In the BIMAP study, 3D micro-PET-CT images of 41 BCa specimens were collected, all from different patients, including 31 NST, 6 ILC, 2 DCIS, and 2 NAC cases. In the BrIMA study, 3D micro-PET-CT images of 152 BCa specimens were collected, once again all from different patients, including 84 NST, 16 ILC, 13 DCIS, 14 NAC, 16 NST-AI, and 9 ILC-AI cases.

6.2.1.2. Micro-PET-CT images of breast cancer specimens

In both studies, we collected 3D micro-PET-CT images of the entire lumpectomy specimens immediately after they were resected from the patient. Patients who participated in these studies received a preoperative injection with [^{18}F]fluorodeoxyglucose ([^{18}F]FDG). In the BIMAP study, patients either received 0.8 MBq/kg or 4 MBq/kg [^{18}F]FDG, while in the BrIMA study all patients received 0.8 MBq/kg [^{18}F]FDG. In the BIMAP study, micro-PET-CT images were either acquired with the β -CUBE[®] and X-CUBE[®] (MOLECUBES, Belgium), or the AURA[®]10 (XEOS, Belgium), while in the BrIMA study, all images were acquired with the AURA[®]10. Although designed for different applications, both scanners offer similar image quality with a submillimetre spatial resolution for PET. To obtain a uniform dataset, all images in both studies were reconstructed to emulate a PET acquisition time of 10 min and an injected dose of 0.8 MBq/kg. Micro-PET images were reconstructed using 20 iterations of the maximum likelihood expectation maximisation (MLEM) algorithm, with an energy window of 50% around the 511 keV photopeak and an isotropic voxel size of 400 μm . CT acquisition took 3 min with the X-CUBE[®] and 1 min with the AURA[®]10. CT images were reconstructed using the image space

reconstruction algorithm (ISRA) (X-CUBE®) or filtered backprojection (FBP) (AURA®10), with an isotropic voxel size of 100 μm . The results were co-registered micro-PET and micro-CT images in DICOM format. 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 [92]. All micro-PET images were corrected for attenuation and median filtered with a 2 mm isotropic kernel to remove noise.

6.2.1.3. Margin status classification labels

At the histopathology department, all lumpectomy specimens were processed following standard-of-care procedures. In the BIMAP dataset, the maximal tumour diameter measured during histopathological processing of the specimens was on average 18.19 ± 9.22 mm, 20.17 ± 8.80 mm, 25.00 ± 15.56 mm, and 6.00 ± 8.49 mm for the NST, ILC, DCIS, and NAC cases. In the BrIMA dataset, the maximal tumour diameter measured during processing of the specimens was on average 19.20 ± 7.04 mm, 18.44 ± 12.97 mm, 23.92 ± 22.24 mm, 15.79 ± 8.94 mm, 18.56 ± 6.09 mm, and 22.56 ± 12.16 mm for the NST, ILC, DCIS, NAC, NST-AI, and ILC-AI cases. The true margin status of all specimens was assessed following ASCO-CAP protocols [176, 177]. The BIMAP dataset was previously described by De Crem et al. [32]. In accordance with De Crem et al. [32], for the BIMAP data, we used the same “ink on tumour” definition of a positive margin for all BCa types, and in case of invasive tumours (NST and ILC) we only accounted for the invasive tumour component. The same definitions were applied to the BrIMA data, except that for DCIS cases, the margin was considered positive when the minimal distance between the tumour and the resection border was ≤ 2 mm. This margin status is further indicated as the “true histopathological margin status” of the specimens. Following these definitions, in the BIMAP dataset, 9 NST, 2 ILC, 1 DCIS, and 1 NAC case were found to have a positive margin. In the BrIMA dataset, 14 NST, 6 ILC, 10 DCIS, 2 NST-AI, and 3 ILC-AI cases were found to have a positive margin. None of the NAC cases in the BrIMA study had a positive margin.

6.2.2. Workflow for margin status prediction

In the previous chapter, DL and intensity thresholding models were developed to segment breast tumours from 2D micro-PET-CT images of breast lamellas. An overview of all trained ResU-Nets and intensity thresholds was provided in Table 5.1 of Chapter 5. To predict the margin status of a BCa specimen, it can be evaluated whether the predicted tumour segmentation touches the border of the specimen. In the current chapter, we evaluate the performance of this approach by comparing the predicted with the true histopathological margin status for BCa specimens.

6.2.2.1. Data preprocessing and splits

Before feeding the micro-PET-CT images of the BIMAP and BrIMA datasets to the DL and intensity thresholding models, they had to undergo several preprocessing steps. DICOM images were converted to NUMPY arrays. The 3D micro-PET images were interpolated to the grid of the micro-CT images with a third-order interpolation. The mouse beds (X-CUBE[®]), or specimen containers (AURA[®] 10), were removed from the micro-CT images, after which the images were thresholded using Otsu's method to define a binary tissue mask [160]. 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 of interest.

Specifically for the DL models, the micro-PET-CT images also had to be min-max scaled to a range of [0,1] and clipped. For this scaling, the same lower and upper limits as defined for the train set were used, as discussed in Section 5.2.2.2 of Chapter 5. 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 input images were padded with background pixels up to the smallest size suited to pass through the network.

For 3 patients in the BIMAP dataset, and 22 patients in the BrIMA dataset, we collected both lamella images with the true tumour segmentations, as well as images of the full specimens with the true histopathological margin status. As a consequence there is an overlap between the dataset with segmentation labels used to train the DL and intensity thresholding models, and the dataset with margin status labels used to validate the models' margin assessment performance. To avoid data leakage, it was ensured that models were never trained and validated on data from the same patient.

6.2.2.2. Deep learning model

We used the ensemble of 25 ResU-Nets, trained in Chapter 5, to segment (pre)malignant breast tissue in the micro-PET-CT images of all BCa specimens in the BIMAP and BrIMA datasets. We used the ensemble of models that take the micro-PET-CT images as input, as these models were found to have the best segmentation performance. To predict tumour segmentations for the 3D micro-PET-CT images using the 2D ResU-Nets, first, 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. Afterwards, the final tumour segmentation was obtained by thresholding this probability map at 0.5. For

the 3 BIMAP and 22 BrIMA cases for which data from the same specimen was previously used during training of the ResU-Nets, only 5/25 ResU-Nets could be used to predict the tumour segmentation to avoid data leakage.

For NST specimens, tumour segmentation was performed using the ResU-Nets trained on NST data. For ILC and DCIS specimens, tumours were segmented using the ResU-Nets obtained through transfer learning on combined ILC and DCIS data. No models were trained to segment NAT breast tumours, pretreated with NAC or aromatase inhibitors, as no ground-truth segmentation labels were available for these cases. As such, for NAC, NST-AI, and ILC-AI specimens, tumour segmentation was performed using the ResU-Nets transfer learned on the combined ILC and DCIS data. This choice was made as successful treatment with NAC and aromatase inhibitors has been shown to reduce [^{18}F]FDG uptake, which is similarly lower in ILC and DCIS cases [178, 179].

Some post-processing steps were applied to the predicted tumour segmentations. To increase specificity, only the largest connected component of the tumour segmentation was retained. It was assumed that this step would have minimal influence on the sensitivity as multifocal BCa is often considered as a contraindication for BCS, making lumpectomy specimens with multifocal tumours uncommon [180]. When multifocal breast tumours are identified through preoperative imaging, this post-processing step can be disabled. In addition, we filled all holes in the tumour segmentation as only the tumour contour is of importance for margin assessment.

To derive the margin status from the obtained tumour segmentation, 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 described in the previous section [160]. The predicted margin status was considered to be positive when the minimum distance between the tumour contour and the specimen contour was ≤ 0.1 mm. This was done to account for uncertainty on the exact location of the specimen contour due to the limited spatial resolution of the micro-CT images. Only for the DCIS specimens in the BrIMA dataset, a larger contour distance ≤ 2 mm was used to define a positive margin. Figure 6.1 visualises the entire workflow when using the DL model to predict the margin status from a micro-PET-CT image of a lumpectomy specimen.

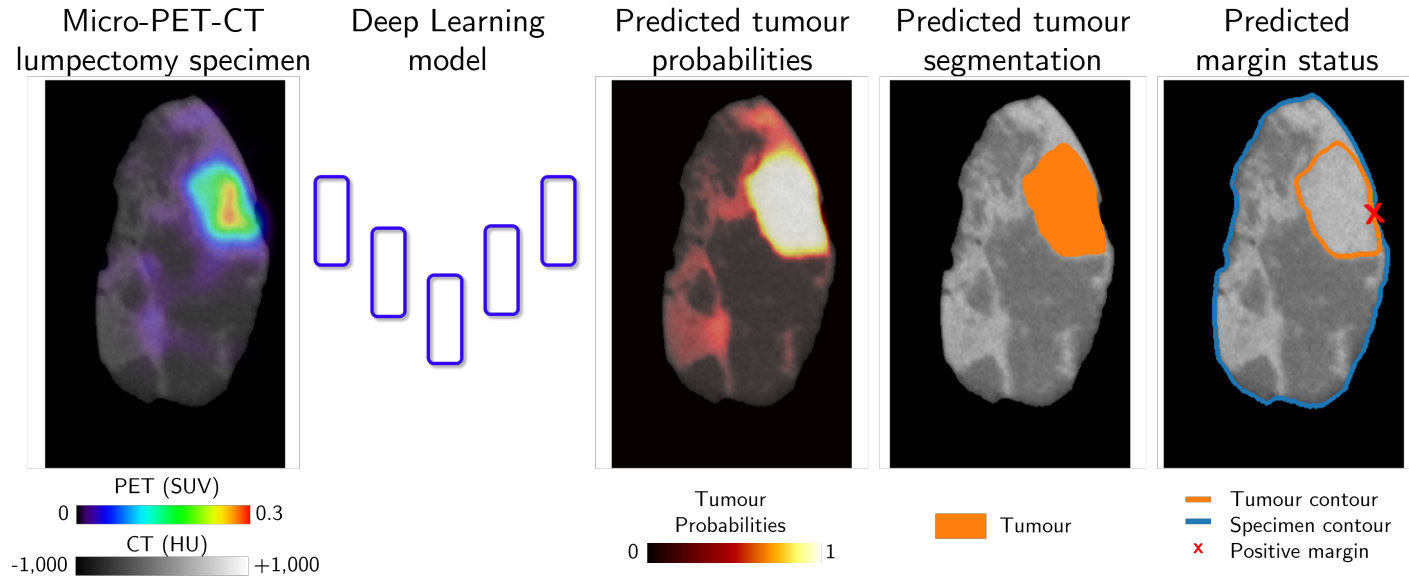


Figure 6.1. Illustration of the automated workflow for tumour segmentation and margin status prediction from a micro-PET-CT image of a lumpectomy specimen using a deep learning (DL) model. The DL model takes a 3D micro-PET-CT image as input, and predicts voxel-wise tumour probabilities ($\in [0,1]$) as output, 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.

6.2.2.3. Intensity thresholding

For margin assessment, we again compared the performance of the DL models with a conventional intensity thresholding approach. We used the obtained ensemble of five intensity thresholds from Chapter 5 to segment (pre)malignant breast tissue in all 3D micro-PET-CT images of the BIMAP and BrIMA datasets. We used the intensity thresholds based on the combined micro-PET-CT inputs, as they led to the best segmentation performance. Accordingly, the 3D micro-PET image was first masked to include only high-density tissues as defined on the 3D micro-CT image. Afterwards, the five relative SUV thresholds were applied to the masked 3D micro-PET image to generate five tumour segmentations. These segmentations were averaged per case and thresholded at 0.5 to obtain the final tumour segmentation. For the 3 BIMAP and 22 BrIMA cases for which data from the same specimen was previously used during training of the intensity thresholds, only 1/5 intensity thresholds could be used to predict the tumour segmentation to avoid data leakage.

In line with the approach for the ResU-Nets, tumour segmentation for NST specimens was performed using the optimal intensity thresholds found for the NST training data. For ILC and DCIS specimens, tumours were segmented using the intensity thresholds found for the combined ILC and DCIS training data. These last intensity thresholds were also used to segment breast tumours in NAC, NST-AI, and ILC-AI specimens. Once again, to increase specificity, only the largest connected component of the predicted tumour segmentation was retained, and hole-filling was performed. The predicted margin status was again considered positive when the minimum distance between the contour of the predicted tumour segmentation and the specimen contour was ≤ 0.1 mm, or ≤ 2 mm for the DCIS specimens in the BrIMA dataset.

6.2.3. Margin assessment performance evaluation

To evaluate the performance of the ResU-Nets and intensity thresholding models in terms of margin assessment accuracy, we compared the predicted margin status with the true histopathological margin status for all BIMAP and BrIMA cases. In addition, the margin status predictions of the ResU-Nets and intensity thresholding models were compared with the predictions of physicians who manually interpreted the same images. As described in Section 2.5.4 of Chapter 2, in the BIMAP study, seven physicians (three breast surgeons, one gynaecology resident, two nuclear medicine physicians, and one radiologist) postoperatively interpreted the micro-PET-CT images and predicted the margin status for all cases. Results of this interpretation study were published by De Crem et al. [32]. We compared the margin status predictions of the ResU-Nets and intensity thresholding models with the postoperative

majority vote of these seven physicians. It is important to note that the metric values reported in Table 2.2 of Chapter 2 evaluate the average margin assessment performance of these seven physicians, and therefore do not correspond to the metrics reported in this chapter to evaluate the performance of the majority vote of these physicians. In the BrIMA study, all micro-PET-CT images were interpreted intraoperatively by the breast surgeons [33]. Furthermore, after study completion, a panel of four breast surgeons participating in the BrIMA study was asked to perform a blinded postoperative re-interpretation of all images. The margin status predictions of the ResU-Nets and intensity thresholding models were compared to the intraoperative predictions of the physicians, as well as to the postoperative majority vote of these four physicians.

Contingency tables were constructed for the margin status predictions performed with the ResU-Nets and intensity thresholding models, as well as for the physicians' intraoperative decisions and postoperative majority votes. Separate contingency tables were generated for each study (BIMAP and BrIMA) and for each BCa subtype. Based on these tables, we calculated sensitivity, specificity, precision, and F1 score for the detection of positive margins. Metrics were again calculated per study and per BCa subtype. To statistically assess differences in margin assessment performance between the ResU-Nets, intensity thresholding models, and the physicians' interpretations, 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 subtype. A significance level $\alpha = 0.05$ was used, and a Bonferroni-Holm correction was performed within each BCa subtype to compensate for multiple testing across different pairwise comparisons of margin assessment strategies. For the BIMAP study, the following three pairwise comparisons were performed: (1) Postoperative majority vote of seven physicians vs. ResU-Net prediction, (2) Postoperative majority vote of seven physicians vs. intensity thresholding prediction, (3) ResU-Net prediction vs. intensity thresholding prediction. A Bonferroni-Holm correction with $m = 3$ was performed. For the BrIMA study, the following six pairwise comparisons were performed: (1) Intraoperative prediction of physicians vs. ResU-Net prediction, (2) Intraoperative prediction of physicians vs. intensity thresholding prediction, (3) Postoperative majority vote of four physicians vs. ResU-Net prediction, (4) Postoperative majority vote of four physicians vs. intensity thresholding prediction, (5) ResU-Net prediction vs. intensity thresholding prediction, and (6) Intraoperative prediction of physicians vs. postoperative majority vote of four physicians. A Bonferroni-Holm correction with $m = 6$ was performed.

Finally, as in any medical classification problem, there is a trade-off between sensitivity, specificity, and precision when performing margin assessment. An additional analysis was performed to get more insight into this trade-off and study the general power of the ResU-Nets and intensity thresholding models to distinguish between positive and negative resection margins. Initially, the margin assessment performance of the ResU-Nets and intensity thresholding models was evaluated at the operating point corresponding to optimal segmentation performance as measured by the Dice similarity coefficient (DSC), hereafter referred to as the “segmentation operating point”. However, selecting an operating point solely based on segmentation performance may not be optimal for margin status classification and alternative operating points should be evaluated. For this purpose, we constructed the receiver operating characteristic (ROC) and precision-recall (PR) curves for both methods. To construct the ROC- and PR-curves for the ResU-Nets, the predicted tumour probability maps were thresholded at different values ranging from 0% to 100% in steps of 5% to obtain tumour segmentations and derived margin status predictions. To construct the curves for the intensity thresholding models, the micro-PET images, masked by the micro-CT segmentations of non-adipose tissue, were thresholded at values ranging from 0% to 100% of the patient-specific maximum PET intensity in 5% steps to obtain tumour segmentations and derived margin status predictions. Separate curves were drawn for each study (BIMAP and BrIMA), as well as for the group of NST tumours and the combined group of all BCa subtypes. The area under the curve (AUC) was calculated for all curves to quantify the overall ability of both methods to discriminate between positive and negative margins.

In addition, we assessed the performance of the ResU-Nets and intensity thresholding models at an alternative operating point selected from the ROC- and PR-curves through a leave-one-out (LOO) strategy. The aim was to match the sensitivity obtained by the physicians. For each case in the BIMAP and BrIMA datasets, the ROC- and PR-curves were constructed using all remaining cases in the dataset. For each case, the operating point was then selected as the highest threshold for which the sensitivity was at least equal to that achieved by the physicians. This procedure was performed for both the ResU-Nets and intensity thresholding models to choose an appropriate tumour probability threshold and a relative PET intensity threshold, respectively. Finally, for each case, the found LOO thresholds were applied to segment the tumour and predict the margin status. The margin assessment performance of the ResU-Nets and intensity thresholding models at this alternative operating point was studied in the same way as before, by constructing the contingency tables, deriving performance metrics, and performing statistical analyses using McNemar’s test.

6.3. Results

We used the trained ResU-Nets and intensity thresholds described in Chapter 5 to segment the tumour and predict the margin status for all 3D micro-PET-CT lumpectomy images in the BIMAP and BrIMA datasets. 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 on an NVIDIA V100 GPU with 32 GB VRAM. Tumour segmentation using five intensity thresholds took only several seconds per 3D image. In this section, we evaluate the margin assessment performance of the ResU-Nets and intensity thresholding models separately for the BIMAP and BrIMA datasets.

6.3.1. Margin assessment performance

6.3.1.1. BIMAP

Figure 6.2 visualises the tumour contours predicted with the ResU-Nets and intensity thresholding models for three representative NST specimens of the BIMAP dataset. Visualisations of predicted tumour contours for two ILC, two DCIS, and two NAC specimens can be found in Figures E.1, E.2, and E.3 of Appendix E, respectively. Contingency tables, comparing the predicted and true histopathological margin status, were created for all BIMAP cases. Separate tables were constructed for the predictions obtained using the ResU-Nets and intensity thresholding models, as well as for the predictions based on the postoperative majority vote of seven physicians. Figure 6.3 displays the contingency results for the NST cases, as well as for all combined BCa subtypes in the BIMAP dataset. Separate contingency tables for the ILC, DCIS, and NAC cases can be found in Figure E.4 of Appendix E. From these contingency tables, sensitivity, specificity, precision, and F1 score for positive margin detection were calculated for each margin assessment strategy. Table 6.1 shows all metric values for the NST cases, as well as for all combined BCa subtypes of the BIMAP dataset. Separate scores for the ILC, DCIS, and NAC cases are shown in Table E.1 of Appendix E. It was confirmed that 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.

P-values resulting from McNemar's test to evaluate differences in margin assessment performance between the ResU-Nets, intensity thresholding models, and the postoperative majority vote of seven physicians, are reported in Figure 6.4 for the NST cases, as well as for all combined BCa subtypes in the BIMAP dataset. Figure E.5 of Appendix E shows separate

results for the ILC, DCIS, and NAC cases. For the NST cases, the margin status predictions obtained with the ResU-Nets were the same as those obtained through postoperative 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. When looking at the combined group of all BCa subtypes in the BIMAP dataset, the predictions obtained with the ResU-Nets were found to be statistically equivalent to the postoperative majority vote of seven physicians. The intensity thresholding models again showed significantly lower performance for margin assessment in the combined group of all BCa subtypes, compared to both the ResU-Net predictions and the postoperative majority vote of seven physicians. Once again, the intensity thresholding models showed lower values for the specificity, precision, and F1 score. When studying the ILC, DCIS, and NAC cases separately, all three margin assessment strategies were found to be statistically equivalent.

Table 6.1. Metric values to evaluate the agreement between the true histopathological and predicted margin status for the BIMAP dataset, shown separately for invasive carcinomas of no special type (NST), and for all breast cancer (BCa) subtypes combined. 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 postoperative majority vote of seven physicians. The performance of these seven physicians was previously described by De Crem et al. [32].

BCa type:		NST (N=31)	
Metrics	Intens thresh	ResU-Net	7 Physicians Postoperative
Sensitivity (Recall)	1.00	0.89	0.89
Specificity	0.59	0.91	0.91
Precision	0.50	0.80	0.80
F1 score	0.67	0.84	0.84
BCa type:		All types (N=41)	
Metrics	Intens thresh	ResU-Net	7 Physicians Postoperative
Sensitivity (Recall)	0.92	0.85	0.77
Specificity	0.54	0.82	0.86
Precision	0.48	0.69	0.71
F1 score	0.63	0.76	0.74

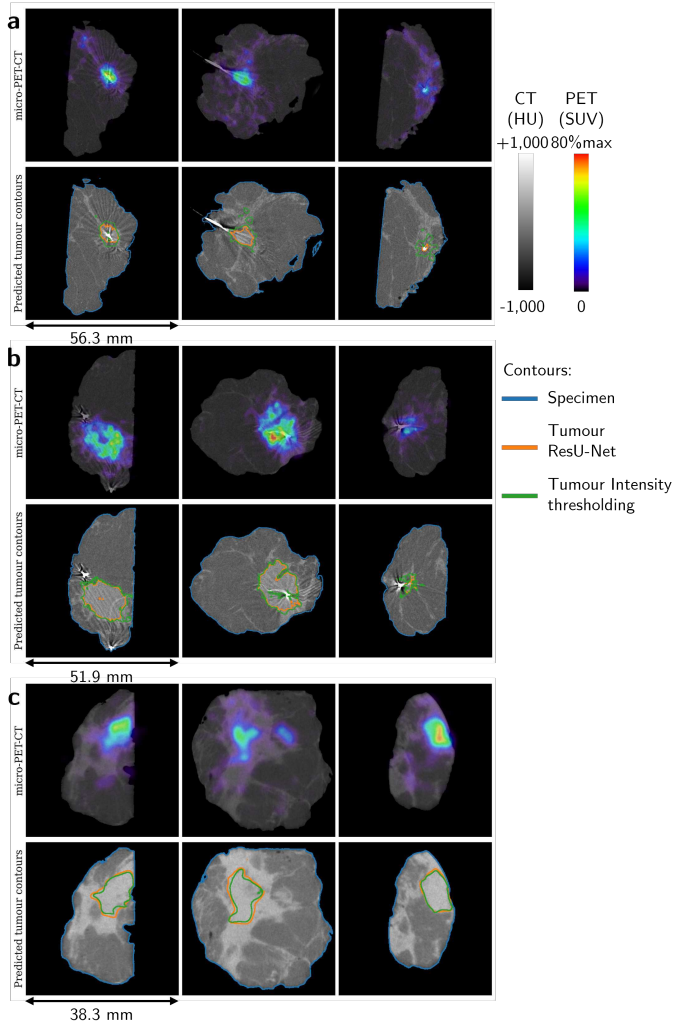


Figure 6.2. Predicted tumour contours for three (a, b, and c) invasive carcinomas of no special type from the BIMAP dataset. For every specimen, three orthogonal slices are visualised. The top row shows the micro-PET-CT input (expressed in standardised uptake values (SUVs) and Hounsfield units (HUs)), and the bottom row shows the contours overlaid on the micro-CT (blue = specimen, orange = tumour predicted by the Residual U-Net (ResU-Net), green = tumour predicted through intensity thresholding). (a) The ResU-Net prediction is true negative, while the intensity thresholding prediction is false positive. (b) The ResU-Net prediction is false negative, while the intensity thresholding prediction is true positive. (c) Both ResU-Net and intensity thresholding predictions are true positive.

		Prediction							
		+	-	Intens thresh		ResU-Net		7 Physicians Postoperative	
Ground Truth	+	TP	FN	9	0	8	1	8	1
	-	FP	TN	9	13	2	20	2	20

	NST (N=31)							
			Intens thresh		ResU-Net		7 Physicians Postoperative	
All types (N=41)	12	1	11	2	10	3		
	13	15	5	23	4	24		

Figure 6.3. Contingency tables for margin status prediction in the BIMAP dataset, shown separately for invasive carcinomas of no special type (NST), and for all breast cancer (BCa) subtypes combined. Margin status was predicted using intensity thresholding (Intens thresh) and the Residual U-Net (ResU-Net) applied to the micro-PET-CT input, as well as through the postoperative majority vote of seven physicians. The predicted margin status was evaluated against the true histopathological margin status. The performance of the seven physicians was previously described by De Crem et al. [32].

NST (N=31)			All types (N=41)		
	Intens thresh	ResU-Net	7 Physicians Postop		
Intens thresh		0.0078	0.0078		
ResU-Net			1.0000		
7 Physicians Postop					

NST (N=31)			All types (N=41)		
	Intens thresh	ResU-Net	7 Physicians Postop		
Intens thresh		0.0039	0.0010		
ResU-Net			0.5000		
7 Physicians Postop					

Figure 6.4. P-values (p) resulting from McNemar’s test to compare different approaches to predict the margin status for the BIMAP dataset, shown separately for invasive carcinomas of no special type (NST), and for all breast cancer (BCa) subtypes combined. The following strategies were compared: postoperative majority voting of seven physicians, prediction with Residual U-Nets (ResU-Nets), and prediction with intensity thresholding (Intens thresh) models. Significant performance differences following Bonferroni-Holm correction ($\alpha = 0.05$, $m = 3$) are indicated in bold.

6.3.1.2. BrIMA

Figure 6.5 visualises the tumour contours predicted with the ResU-Nets and intensity thresholding models for three representative NST specimens of the BrIMA dataset. Visualisations of predicted tumour contours for two ILC, two DCIS, and three NAT specimens (one NST after NAC, one NST-AI, and one ILC-AI) can be found in Figures E.6, E.7, and E.8 of Appendix E, respectively. Contingency tables were created for all BrIMA cases. Separate tables were constructed for the predictions obtained using the ResU-Nets and intensity thresholding models, as well as for the physicians' intraoperative predictions and the postoperative majority vote of four physicians. Figure 6.6 displays the contingency results for the NST cases, as well as for all combined BCa subtypes in the BrIMA dataset. Separate contingency tables for the ILC, DCIS, and combined NAT cases can be found in Figure E.9 of Appendix E. From these contingency tables, sensitivity, specificity, precision, and F1 score for positive margin detection were calculated for each margin assessment strategy. Table 6.2 shows all metric values for the NST cases, as well as for all combined BCa subtypes of the BrIMA dataset. Separate scores for the ILC, DCIS, and combined NAT cases are shown in Table E.2 of Appendix E.

P-values resulting from McNemar's test to evaluate differences in margin assessment performance between the ResU-Nets, intensity thresholding models, the intraoperative physicians' predictions, and the majority vote of four physicians, are reported in Figure 6.7 for the NST cases, as well as for all combined BCa subtypes in the BrIMA dataset. Figure E.10 of Appendix E shows separate results for the ILC, DCIS, and combined NAT cases. For the NST cases, the margin assessment performance of the ResU-Nets was found to be statistically equivalent to the intraoperative physicians' predictions, as well as to the postoperative majority vote of four physicians. While the ResU-Nets showed a higher sensitivity, the specificity was lower compared to the physicians' predictions. The intensity thresholding models showed significantly lower performance for the NST cases, compared to both the ResU-Net predictions and the intraoperative and postoperative physicians' predictions. The intensity thresholding models showed lower values for the specificity, precision, and F1 score. When looking at the combined group of all BCa subtypes in the BrIMA dataset, the same trends were found. The performance of the ResU-Nets was again statistically equivalent to the intraoperative and postoperative physicians' predictions. The intensity thresholding models again showed significantly lower performance for margin assessment in the combined group of all BCa subtypes, compared to both the ResU-Net predictions and the intraoperative and postoperative physicians' predictions, with lower values for the specificity, precision, and F1 score. When studying the ILC, DCIS, and NAT cases separately, it was found that for the DCIS cases, none of the differences between the four margin assessment strategies was statistically significant. For the ILC and combined NAT cases, the

ResU-Net predictions were found to be statistically equivalent to the intraoperative physicians' predictions and the postoperative majority vote of four physicians. The performance of the intensity thresholding models for the ILC and combined NAT cases was found to be significantly different from the performance of the ResU-Nets and the intraoperative predictions of the physicians, with a lower specificity, precision, and F1 score.

Table 6.2. Metric values to evaluate the agreement between the true histopathological and predicted margin status for the BrIMA dataset, shown separately for invasive carcinomas of no special type (NST), and for all breast cancer (BCa) subtypes combined. 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 intraoperative interpretations of the physicians, as well as to the postoperative majority vote of four physicians. The performance of the physicians is also described by Göker et al. [33].

BCa type:		NST (N=84)		
Metrics	Intens thresh	ResU-Net	Physicians Intraoperative	4 Physicians Postoperative
Sensitivity (Recall)	1.00	0.86	0.71	0.71
Specificity	0.14	0.54	0.66	0.67
Precision	0.19	0.27	0.29	0.30
F1 score	0.32	0.41	0.42	0.43
BCa type:		All types (N=152)		
Metrics	Intens thresh	ResU-Net	Physicians Intraoperative	4 Physicians Postoperative
Sensitivity (Recall)	1.00	0.83	0.77	0.83
Specificity	0.15	0.56	0.67	0.55
Precision	0.26	0.36	0.41	0.35
F1 score	0.41	0.50	0.53	0.50

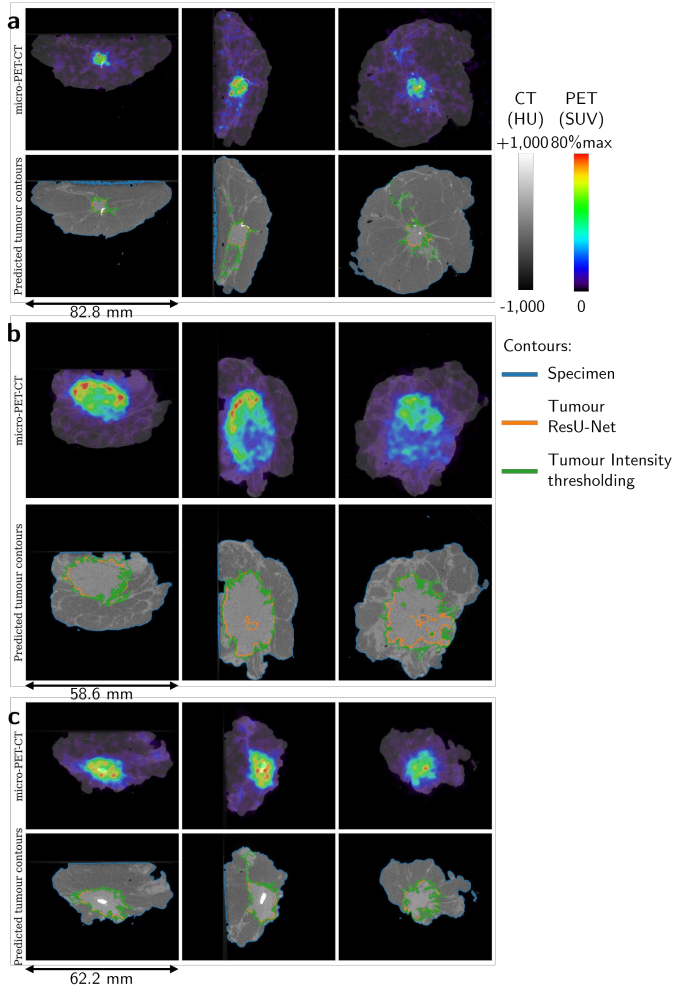


Figure 6.5. Predicted tumour contours for three (a, b, and c) invasive carcinomas of no special type from the BrIMA dataset. For every specimen, three orthogonal slices are visualised. The top row shows the micro-PET-CT input (expressed in standardised uptake values (SUVs) and Hounsfield units (HUs)), and the bottom row shows the contours overlaid on the micro-CT (blue = specimen, orange = tumour predicted by the Residual U-Net (ResU-Net), green = tumour predicted through intensity thresholding). (a) The ResU-Net prediction is true negative, while the intensity thresholding prediction is false positive. (b) The ResU-Net prediction is false negative, while the intensity thresholding prediction is true positive. (c) Both ResU-Net and intensity thresholding predictions are true positive.

		Prediction			Intens thresh		ResU-Net		Physicians		4 Physicians	
		+	-						Intraoperative	Postoperative		
Ground Truth	+	TP	FN	NST (N=84)	14	0	12	2	10	4	10	4
	-	FP	TN		60	10	32	38	24	46	23	47
					35	0	29	6	27	8	29	6
					99	18	51	66	39	78	53	64
					All types (N=152)							

Figure 6.6. Contingency tables for margin status prediction in the BrIMA dataset, shown separately for invasive carcinomas of no special type (NST), and for all breast cancer (BCa) subtypes combined. Margin status was predicted using intensity thresholding (Intens thresh) and the Residual U-Net (ResU-Net) applied to the micro-PET-CT input, as well as through intraoperative physicians' assessments and the postoperative majority vote of four physicians. The predicted margin status was evaluated against the true histopathological margin status. The performance of the physicians is also described by Göker et al. [33].

NST (N=84)					All types (N=152)				
	Intens thresh	ResU-Net	Physicians Intraop	4 Physicians Postop		Intens thresh	ResU-Net	Physicians Intraop	4 Physicians Postop
Intens thresh		0.0000	0.0000	0.0000			0.0000	0.0000	0.0000
ResU-Net			0.0755	0.0192				0.0336	0.8679
Physicians Intraop				1.0000					0.0259
4 Physicians Postop									

Figure 6.7. P-values (p) resulting from McNemar's test comparing different approaches to predict the margin status for the BrIMA dataset, shown separately for invasive carcinomas of no special type (NST), and for all breast cancer (BCa) subtypes combined. Following strategies were compared: intraoperative interpretations of physicians, postoperative majority voting of four physicians, prediction with Residual U-Nets (ResU-Nets), and with intensity thresholding (Intens thresh) models. Significant differences following Bonferroni-Holm correction ($\alpha = 0.05$, $m = 6$) are shown in bold.

6.3.2. Performance stability across operating points

Both for the BIMAP and BrIMA datasets, the margin assessment performance of the ResU-Nets was found to be significantly better than that of the intensity thresholding models, both for the NST tumours and for the combined group of all BCa subtypes. However, inspection of the ROC- and PR-curves in Figures 6.8 and 6.9, indicates that operating points exist at which the performance of the ResU-Nets and intensity thresholding models is more comparable. Figures 6.8 and 6.9 show the ROC- and PR-curves for margin status prediction with the ResU-Nets and intensity thresholding models for the BIMAP and BrIMA datasets, respectively. Separate curves are drawn for the NST cases and for all combined BCa subtypes in the BIMAP and BrIMA datasets. This observation suggests that the operating point previously selected for the intensity thresholding models may not be optimal for margin assessment. In Section 6.3.1, we evaluated the performance of both models at the “segmentation operating point”, corresponding to optimal tumour segmentation performance as measured by the DSC. A potential mismatch may exist between the intensity thresholds that optimise tumour segmentation performance, and those that maximise margin assessment performance. One possible explanation is that the DSC does not explicitly focus on the correct delineation of the tumour contour, which is critical for an accurate margin status prediction. Another possibility is that the dataset with segmentation labels was too small to choose a robust intensity threshold for tumour segmentation.

Because of these considerations, an alternative strategy was explored to select an operating point on the ROC- and PR-curves. We assessed the performance of the ResU-Nets and intensity thresholding models at an operating point chosen through a LOO strategy, trying to match the sensitivity achieved by the physicians. This procedure was performed for both the ResU-Nets and intensity thresholding models. All results of this analysis can be found in Appendix F, both for the BIMAP and BrIMA datasets. Margin assessment performance of the ResU-Nets and intensity thresholding models at this operating point was calculated for the NST cases and for all combined BCa subtypes in both datasets. Figures F.1 and F.3 show the contingency tables for the BIMAP and BrIMA datasets, respectively. The sensitivity, specificity, precision, and F1 score for positive margin detection can be found in Tables F.1 and F.2 for the BIMAP and BrIMA datasets, respectively. P-values resulting from McNemar’s test to evaluate differences in margin assessment performance between the ResU-Nets, the intensity thresholding models, and the intraoperative and postoperative physicians’ predictions are reported in Figures F.2 and F.4 for the BIMAP and BrIMA datasets, respectively.

In accordance with what was reported for the segmentation operating point, it was found that the performance of the ResU-Nets was still

statistically equivalent to the intraoperative physicians' predictions and to the postoperative majority vote of the physicians, both for the NST cases and the combined group of all BCa subtypes, and both for the BIMAP and BrIMA datasets. In contrast to the findings reported for the segmentation operating point, for the BIMAP dataset, no statistically significant differences in performance were observed between the intensity thresholding models and either the ResU-Nets or the postoperative majority vote of seven physicians, both for the NST cases and the combined group of all BCa subtypes. However, for the BrIMA dataset, the performance of the intensity thresholding models was again found to be significantly different from the ResU-Nets, the intraoperative physicians' predictions, and the postoperative majority vote of four physicians, both for the NST cases and the combined group of all BCa subtypes. The intensity thresholding models obtained a lower specificity, precision, and F1 score compared to the physicians' predictions.

The persistently lower performance of the intensity thresholding models compared to the ResU-Nets for the BrIMA dataset, indicates that identifying a robust PET intensity threshold that leads to correct margin status predictions across a large number of cases remains challenging. Looking at the ROC- and PR-curves for the intensity thresholding models, it can be noted that they evolve in large steps. A relatively small change of 5% in the intensity threshold can lead to strong variations in the performance metrics. Although performing a more fine-grained search for a suited threshold might improve the final performance metrics, the need for such a fine-grained search also suggests instability and sensitivity to the data at hand. In contrast, looking at the ROC- and PR-curves for the ResU-Nets, it can be noted that around the performance level of the physicians, multiple operating points with a similar performance can be found, indicating more stability. Overall, these findings suggest that intensity thresholding is less robust and less reliable than the ResU-Nets for margin status prediction. Finally, many alternative operating points could be targeted, depending on the relative importance assigned to reaching a high sensitivity, specificity, or precision.

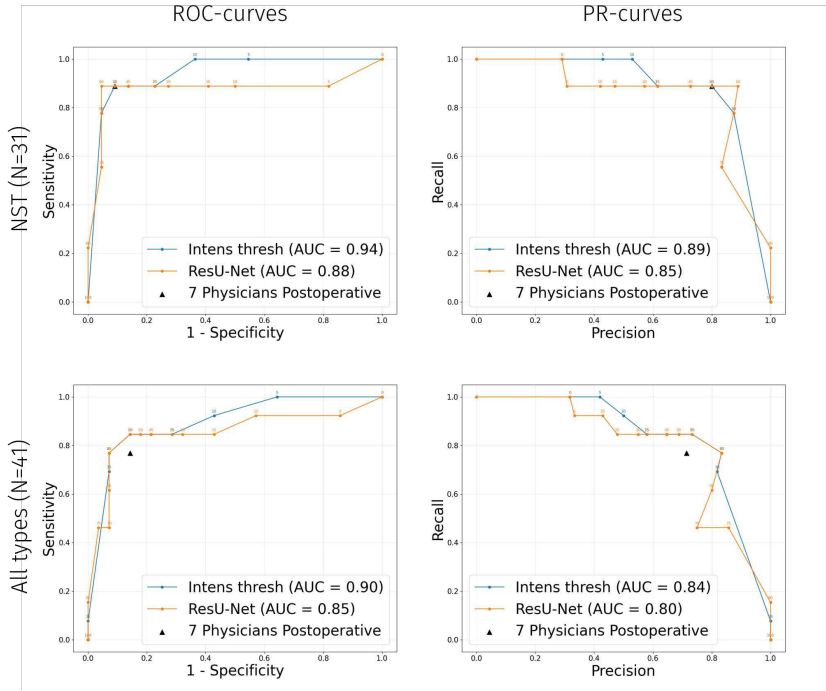


Figure 6.8. Receiver operating characteristic (ROC) and precision-recall (PR) curves for margin status prediction for the BIMAP dataset, shown separately for invasive carcinomas of no special type (NST), and for the combined group of all breast cancer (BCa) subtypes. For intensity thresholding (Intens thresh), the micro-PET signal, masked by the micro-CT segmentation of non-adipose tissue, was thresholded at values ranging from 0% to 100% of the patient-specific maximum PET intensity in 5% steps to predict the margin status. For the Residual U-Nets (ResU-Nets), tumour probability maps were predicted from the micro-PET-CT input, and thresholded at values ranging from 0% to 100% in 5% steps to predict the margin status. The area under the curve (AUC) values are shown in the figure legends. All figures also display the performance of the postoperative majority vote of seven physicians. The performance of these seven physicians was previously described by De Crem et al. [32].

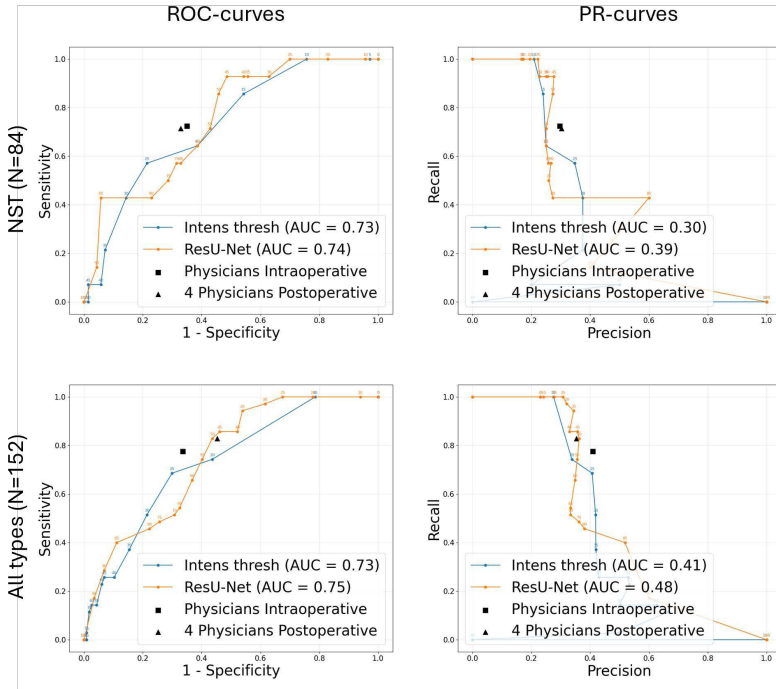


Figure 6.9. Receiver operating characteristic (ROC) and precision-recall (PR) curves for margin status prediction for the BrIMA dataset, shown separately for invasive carcinomas of no special type (NST), and for the combined group of all breast cancer (BCa) subtypes. For intensity thresholding (Intens thresh), the micro-PET signal, masked by the micro-CT segmentation of non-adipose tissue, was thresholded at values ranging from 0% to 100% of the patient-specific maximum PET intensity in 5% steps to predict the margin status. For the Residual U-Nets (ResU-Nets), tumour probability maps were predicted from the micro-PET-CT input, and thresholded at values ranging from 0% to 100% in 5% steps to predict the margin status. The area under the curve (AUC) values are shown in the figure legends. All figures also display the performance of the intraoperative interpretations by the physicians, and the postoperative majority vote of four physicians. The performance of the physicians is also described by Göker et al. [33].

6.4. Discussion

To our knowledge, we were the first to apply the well-known ResU-Net model to perform automatic tumour segmentation and margin assessment based on micro-PET-CT images of lumpectomy specimens. In this chapter, we used the BIMAP and BrIMA datasets, consisting of 3D micro-PET-CT lumpectomy images with true histopathological margin status labels, to evaluate the performance of the ResU-Nets trained in Chapter 5 in terms of margin assessment. When focussing on the results in Section 6.3.1, it was found that for the NST tumours, the ResU-Nets reached F1 scores of 84% and 41% for the BIMAP and BrIMA datasets, respectively. For the BIMAP dataset, the ResU-Nets were found to match the postoperative majority vote of seven physicians. For the BrIMA dataset, the performance of the ResU-Nets was also found to be statistically equivalent to the intraoperative physicians' assessments and to the postoperative majority vote of four physicians. For the NST cases, the ResU-Nets were found to outperform intensity thresholding models both for the BIMAP and BrIMA datasets. For all BCa subtypes combined, the ResU-Nets reached F1 scores of 76% and 50% for the BIMAP and BrIMA datasets, respectively. Both for the BIMAP and BrIMA datasets, the performance of the ResU-Nets was again found to be statistically equivalent to the intraoperative and postoperative predictions of the physicians, while outperforming intensity thresholding. This finding is particularly relevant, as the combined cohort of all BCa subtypes reflects the broader BCa population, with subtype proportions approximately corresponding to their prevalence in larger clinical populations, as discussed in Section 2.2.4.3 of Chapter 2.

While similar trends were observed when studying the ILC, DCIS, and NAT cases separately, the more limited datasets for these subgroups did not always allow us to show statistically significant differences in margin assessment performance between the ResU-Nets and intensity thresholding models. The performance of the ResU-Nets and intensity thresholding models were only found to be significantly different for the ILC and NAT cases in the BrIMA dataset. The performance of the intensity thresholding models was also shown to be significantly different from the intraoperative physicians' predictions for the ILC and NAT cases in the BrIMA dataset. For these cases, the intensity thresholding models showed a lower specificity, precision, and F1 score compared to the ResU-Nets and intraoperative and postoperative physicians' predictions. While the segmentation models were only trained on NST, ILC, and DCIS tumours, it was found that the ResU-Nets could also be used to segment NAT breast tumours and predict the margin status, with a performance similar to physicians interpreting the images.

Overall, the same trends were observed in both the BIMAP and BrIMA datasets, although effects were more pronounced in the BrIMA dataset. For instance, while the intensity thresholding models already showed lower specificity and precision than the ResU-Nets in the BIMAP dataset,

these metrics substantially decreased in the BrIMA dataset. Due to the increased sample sizes in the BrIMA dataset, statistically significant differences could also be shown in subgroups of more rare BCa types. In general, the ResU-Nets achieved higher performance on the BIMAP dataset compared to the BrIMA dataset. However, this pattern was consistent with the performance of the physicians. Importantly, the BrIMA dataset is considered more clinically relevant, as it was collected following a more stringent study design in a larger, more diverse, and more representative patient population. Unlike the BIMAP study, which can be considered more as a pilot study conducted primarily in a research setting, the BrIMA study involved multiple BCa centres across several countries and integrated micro-PET-CT imaging directly into the clinical workflow.

Other groups have used DL to create similar workflows for automatic IMA in BCa based on other intraoperative imaging techniques. Veluponnar et al. [170] and Yeung et al. [171] applied a similar approach to ultrasound (US) images of lumpectomy specimens. As discussed in Chapter 5, they first also developed DL models to segment tumours in their respective lumpectomy images, and afterwards, they 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 (≤ 2 mm) for 86 patients (of which 91% were of the NST type). For the NST tumours in the BIMAP dataset, we found a slightly lower sensitivity of 89%, with a higher specificity of 91%. For the NST tumours in the BrIMA dataset, we found a lower sensitivity and specificity of 86% and 54%, respectively. However, the scores reported by Veluponnar et al. [170] 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 (≤ 2 mm) margins, which does not match the clinical “ink on tumour” definition for invasive carcinomas. Yeung et al. [171] reached a sensitivity of 75% and a specificity of 43% when predicting the margin status for five breast specimens (of unknown BCa subtype). We reached a higher performance for the NST tumours and for the combined group of all BCa subtypes in both the BIMAP and BrIMA datasets. However, the dataset used by Yeung et al. [171] 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. Furthermore, automatic margin assessment based on micro-PET-CT images of cancer specimens has been performed before. Moraitis et al. [35] proposed an iterative thresholding method to segment tumours in [^{18}F]PSMA-1007 (prostate-specific membrane antigen) micro-PET-CT images of resected prostate specimens. 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 [35]. Despite the small dataset, the results look promising. In future work,

we could also investigate the performance of such iterative thresholding method for lumpectomy specimens.

This study contains several limitations related to the data. First, there might be slight inaccuracies in the data labels. Mistakes in the margin status labels might occur due to sampling bias in the histopathological workup [31]. Second, the datasets in this work are still rather small for BCa types other than the most common NST type. Especially for DCIS cases, more data is required to validate the ResU-Nets as currently no significant difference could be shown for the margin assessment performance of the ResU-Nets and intensity thresholding models. As mentioned in Chapter 5, it is assumed that the BCa type is known from preoperative biopsy, such that the most suited tumour segmentation model can be applied. Performance might slightly decrease when using a model not tailored to the BCa subtype. The model might also incorrectly detect tumours in entirely benign specimens, as in this case, relative scaling might amplify noise in the micro-PET images. Third, the ResU-Nets trained for breast tumour segmentation are 2D models, while the goal is to segment 3D tumours. As mentioned in Chapter 5, this means that the 3D information in the micro-PET-CT images cannot be exploited by the 2D ResU-Nets, possibly limiting the performance. Moreover, mismatches between the 2D lamella images used to train the ResU-Nets, 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 these two datasets, 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 this DL-based software could be integrated in the micro-PET-CT specimen scanners and applied in clinical practice, some additional steps are required. Additional data still needs to be collected to further improve and validate the DL models, especially for BCa types other than the most common NST type. It may also be valuable to examine whether the proposed models correctly predict the orientation of positive margins. In addition, it should be assessed how we can quantify the model's confidence when predicting the margin status. In the next chapter, we will calibrate the predicted tumour probability maps, which express spatially-resolved prediction uncertainty, so that the pixel-level probabilities reflect the true likelihood of malignancy, thereby enhancing model trustworthiness. Finally, we should also investigate how the DL models can support physicians to improve performance and efficiency for micro-PET-CT-based IMA in BCS.

6.5. Conclusions

We validated the ability of ResU-Nets to segment the tumour and predict the margin status based on 3D micro-PET-CT images of lumpectomy specimens. The ResU-Nets showed margin assessment performance statistically equivalent to intraoperative and postoperative predictions of physicians, while consistently outperforming intensity thresholding for NST tumours and for the combined group of BCa subtypes in the BIMAP and BrIMA datasets. Although statistical significance could not always be shown for smaller ILC, DCIS, and NAT subsets, the ResU-Nets generally showed higher specificity, precision, and F1 scores compared to intensity thresholding models. Notably, despite not being trained on NAT tumours, the model achieved performance comparable to physicians for this subtype, highlighting its robustness. The fact that the fully automatic ResU-Net showed a margin assessment performance comparable to physicians, highlights its potential as a clinical support tool, 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, capable of predicting both tumour segmentation and margin status, represents an important step towards a decision-support system that enhances micro-PET-CT-based IMA in BCa, facilitating the adoption of this technique in clinical practice.

7 | Calibration of tumour probability maps for trustworthy interpretation of micro-PET-CT images of lumpectomy specimens

Abstract

Introduction: Accurate interpretation of high-resolution (micro-) positron emission tomography (PET) images, in combination with computed tomography (CT) images, of lumpectomy specimens for intraoperative margin assessment during breast-conserving surgery remains challenging in some cases. A decision-support system is therefore being developed to aid physicians in image interpretation. Previous work (Chapters 5 and 6) introduced an ensemble of Residual U-Nets (ResU-Nets) for breast tumour segmentation and margin status prediction. While performance looked promising, reliable clinical use of such deep learning (DL) models also requires insight into prediction uncertainty, which is absent in current binary outputs. In this study, we aim to improve model trustworthiness by calibrating predicted tumour probability maps, making the spatially-resolved uncertainty they express more reliable to support risk-informed clinical decisions.

Methods: Tumour probability maps predicted by the ResU-Nets were calibrated using the inductive cross-Venn-ABERS method. Calibration was performed and quantitatively evaluated using a dataset of 53 2D micro-PET-CT images from lumpectomy lamellas of 19 patients with early-stage breast cancer of no special type (NST), along with ground-truth histopathology-derived tumour segmentations. Calibrated tumour probability maps were also predicted for 84 3D micro-PET-CT images of NST lumpectomy specimens. The impact of calibration on tumour segmentation and margin assessment performance of the models was evaluated using the lamella and full-specimen datasets, respectively. Additionally, calibrated tumour probability maps were qualitatively analysed for both lamella and full-specimen data.

Results: Venn-ABERS substantially improved the calibration of pixel-level tumour probabilities, reducing expected calibration error, binary cross-entropy loss, and Brier loss by 60.5%, 17.7%, and 14.8%, to final values of 0.017, 0.172, and 0.052, respectively. Tumour segmentation performance remained stable, while margin assessment showed a slight decrease in area under the curve of the precision-recall curve. Qualitatively, the probability maps provided transparent, spatially-resolved uncertainty visualisations, and calibration made these estimates more reliable, allowing uncertain tumour boundaries and ambiguous margin regions to be identified.

Conclusions: Calibration using Venn-ABERS improved alignment between predicted pixel-level tumour probabilities and observed frequencies of malignant pixels, while largely preserving segmentation and margin assessment performance. Calibrated tumour probability maps offered transparent insight into spatially-resolved prediction uncertainty, potentially supporting more efficient and informed clinical decisions. While efforts are required to evaluate the interpretability and clinical utility of the calibrated tumour probability maps for margin assessment, this work marks an important step towards trustworthy deployment of the DL models.

7.1. Introduction

Interpreting high-resolution (micro-) positron emission tomography (PET) images, in combination with computed tomography (CT) images, of lumpectomy specimens for intraoperative margin assessment (IMA) during breast-conserving surgery (BCS) can be time-consuming and challenging in some cases. To aid physicians in image interpretation, we are developing a decision-support system. In Chapter 5 we introduced an ensemble of deep learning (DL) models for automatic segmentation of malignant breast tissue in micro-PET-CT images. In Chapter 6, it was demonstrated how the margin status of a lumpectomy specimen could be derived from these predicted tumour segmentations, achieving margin assessment performance statistically equivalent to intraoperative and postoperative predictions of physicians in a multicentre dataset. However, the ultimate goal of such DL models is not merely predicting binary tumour segmentations or margin status labels, but informing clinical decisions. Binary outputs lack transparency on prediction uncertainty, which is essential for assessing prediction reliability and identifying potential errors. For safe clinical integration, models should provide probabilistic outputs that reflect their confidence, thereby improving interpretability and trust.

Uncertainty estimation is increasingly recognised as a key component in bridging the gap between algorithmic predictions and clinical practice [118]. Uncertainty in DL generally arises from two main sources: aleatoric (data-related) and epistemic (model-related) uncertainty [104, 118, 119]. Aleatoric uncertainty stems from inherent randomness and unpredictability of the studied system, such as noise in the data. Epistemic uncertainty arises from a lack of knowledge about the system caused by incomplete training data, or suboptimal model design and optimisation choices. While aleatoric uncertainty is often irreducible, epistemic uncertainty can be reduced given enough data or improvements in model design. Accurately quantifying uncertainty is crucial for identifying unreliable predictions. An important requirement for meaningful and reliable uncertainty estimation is calibration of probabilities, where predicted probabilities reflect the true likelihood or frequency of an outcome [119, 181]. Poor probability calibration, common in DL, often leads to overconfident predictions posing potential clinical risks [119, 181]. Conformal prediction methods, introduced by Vovk et al. [182], offer a promising solution by providing statistically valid uncertainty estimates without strong assumptions on data distribution, dataset size, or underlying DL models [118, 119].

In this chapter, we aim to improve the trustworthiness of our DL models by calibrating their predicted tumour probability maps using methods from the conformal prediction framework. By moving from binary segmentations to probability maps, we provide transparent, spatially-

resolved insight into prediction uncertainty, and by calibrating these maps we aim to make that uncertainty more reliable. Understanding uncertainty in the location of the tumour boundary relative to the specimen margin helps physicians distinguish between confident and ambiguous regions. This may allow more informed, patient-specific decision-making and can improve efficiency by directing attention to uncertain areas. Ultimately, this approach aims to enhance transparency, reliability, and trust, facilitating safe clinical adoption of the DL models.

7.2. Methods

7.2.1. Datasets

7.2.1.1. 2D micro-PET-CT of lamellas with tumour segmentation labels

To calibrate the tumour probability maps predicted by the DL models, we used a dataset of 2D micro-PET-CT slices of lumpectomy lamellas with corresponding true tumour segmentations derived from aligned histopathological whole slide images (WSIs). The co-registration method used to construct this dataset was described in Chapter 4, and the same dataset was previously employed in Chapter 5 to train and evaluate the DL models for tumour segmentation. In this study, we focused on the largest subgroup of invasive carcinomas of no special type (NST). The dataset consists of 53 2D micro-PET-CT images from lumpectomy lamellas of 19 patients with early-stage NST breast cancer (BCa), along with ground-truth tumour annotations. These annotations provide pixel-wise binary tumour labels, required to perform and evaluate the calibration of pixel-level tumour probabilities. Further details on dataset collection including study design, imaging protocol, and annotation procedures, are provided in Section 5.2.1 of Chapter 5.

7.2.1.2. 3D micro-PET-CT of specimens with margin classification labels

To further evaluate the calibration approach, we used a dataset of 3D micro-PET-CT images of lumpectomy specimens with known true histopathological margin status. For this study, we focused on the NST subgroup from the multicentre BrIMA study [33], previously introduced in Section 2.5.4 of Chapter 2, and used to evaluate margin assessment performance of the DL models in Chapter 6. The dataset contains 3D micro-PET-CT images of 84 lumpectomy specimens from patients with early-stage NST BCa, of which 14 cases had a positive margin according to the “ink on tumour” definition. These images with ground-truth margin status labels can be used to assess the value of calibrated tumour

probability maps to aid in margin assessment. Further details on the collection of this dataset are provided in Section 6.2.1 of Chapter 6.

7.2.2. Venn-ABERS calibration of pixel-level tumour probabilities

The ensemble of 25 Residual U-Nets (ResU-Nets), introduced in Chapter 5, predicts tumour probability maps with pixel-level probabilities $\in [0, 1]$ from micro-PET-CT images. Our goal is to calibrate these pixel-level probabilities such that they better reflect the true frequency of malignant pixels. To this end, we applied Venn-ABERS calibration, a post hoc method proposed by Vovk and Petej [183] to calibrate predicted probabilities for binary classification problems. This approach is part of the conformal prediction framework and offers theoretical guarantees of valid calibration, regardless of the data distribution, dataset size, or underlying DL model [183].

In the computationally efficient inductive variant of Venn-ABERS, the available dataset with known classification labels is split into a training and calibration set [119, 184]. The DL model is trained on the training set and subsequently used to predict probability outputs for the calibration set. Isotonic regression can be applied to the calibration set with known true classification labels to learn a monotonically increasing, piecewise constant mapping from uncalibrated to calibrated probabilities. When making a prediction for a new test sample, its predicted uncalibrated probability is temporarily added to the calibration set twice, once assuming the sample belongs to the negative class (label = 0) and once assuming it belongs to the positive class (label = 1). Afterwards, two separate isotonic regressions are performed, resulting in two probabilities for the test sample: p_0 and p_1 . These probabilities form an interval $[p_0, p_1]$, representing lower and upper bounds on the probability of the test sample belonging to the positive class. Vovk et al. [184] showed that this interval is guaranteed to be well calibrated. For practical use, this interval is often reduced to a single probability p using Equation 7.1. Although this combination sacrifices the strict theoretical guarantees, it has been shown to consistently outperform more traditional calibration methods such as Platt scaling and standard isotonic regression in a wide variety of classification problems [119, 183, 184].

$$p = \frac{p_1}{1 - p_0 + p_1} \quad (7.1)$$

In this work, the segmentation task is approached as a pixel-wise classification problem, enabling the application of Venn-ABERS calibration at the pixel level. The ResU-Nets were trained on 2D micro-PET-CT images of lumpectomy lamellas with known tumour segmentations, using a group nested five-fold cross-validation scheme as described in Section 5.2.2.2 of Chapter 5. Given the limited dataset size, we did not retrain the ResU-Nets

and used the validation folds as calibration sets. For each test sample, this resulted in five pixel-level tumour probability intervals $[p_0, p_1]$, one per validation fold. Figure 7.1 illustrates how one interval is obtained for a single test pixel using one validation (or calibration) fold. The five obtained probability intervals were combined into a single calibrated probability p per pixel using the geometric mean (GM) across all five p_0 and p_1 values, as defined in Equation 7.2. This approach to use inductive Venn-ABERS in a cross-validation scheme, also referred to as cross-Venn-ABERS, was introduced by Vovk et al. [184]. Although this variant does not retain formal calibration guarantees, it has shown strong empirical performance [184]. When applying the full ensemble of 25 ResU-Nets to unseen micro-PET-CT images (e.g., specimens from the BrIMA dataset), the same combination strategy can be used, now computing the geometric mean across 25 p_0 and p_1 values. In practice, Venn-ABERS calibration was implemented using the Python library by Petej [185].

$$p = \frac{\text{GM}(p_1)}{\text{GM}(1 - p_0) + \text{GM}(p_1)} \quad (7.2)$$

7.2.3. Calibration evaluation

7.2.3.1. Quantitative calibration assessment

Venn-ABERS calibration was applied to all tumour probability maps predicted for the test cases in the dataset of 2D micro-PET-CT images with ground-truth tumour segmentations. Calibration performance was quantitatively assessed by comparing predicted pixel-level tumour probabilities with their true labels, both before and after application of the Venn-ABERS method. Calibration plots were constructed using all pixels from the test samples [119]. In these plots, pixels are grouped into bins based on their predicted probabilities. For each bin, the observed frequency of true malignant pixels is plotted against the mean predicted tumour probability across all pixels in that bin. A perfectly calibrated model would yield points along the diagonal (45° angle), indicating agreement between the predicted tumour probabilities and observed frequencies of malignant pixels.

Next, we calculated several quantitative calibration metrics: the expected calibration error (ECE), the binary cross-entropy (BCE) loss, and the Brier loss. The ECE, defined in Equation 7.3, quantifies how well predicted probabilities align with the observed frequency of the positive class [119]. To compute it, the N data points (pixels) are partitioned into M equally spaced bins based on their predicted tumour probabilities. For each bin B_m , the accuracy (acc) and confidence (conf) are computed. The accuracy corresponds to the fraction of pixels with a positive label within the bin, with $y_i \in \{0, 1\}$ the true label of pixel i and $|B_m|$ the total number of

pixels in bin B_m . The confidence represents the average predicted tumour probability of all pixels in the bin, with $\hat{p}_i \in [0, 1]$ the predicted tumour probability for pixel i . The ECE is then obtained as the weighted average of the absolute differences between accuracy and confidence across all bins, where the weights reflect the proportion of samples in each bin. This score can also be interpreted as the weighted average distance from the diagonal in the calibration plot. Lower ECE values $\in [0, 1]$ indicate better calibration. The BCE loss or log-loss, given in Equation 7.4, measures how well the predicted tumour probabilities align with the true binary labels by penalising confident but incorrect predictions more strongly through the logarithmic function. Lower values of the positive BCE loss indicate better calibration. The Brier loss, defined in Equation 7.5, measures the mean squared difference between predicted probabilities and true binary labels across all pixels [119]. By squaring the errors, larger deviations are penalised more heavily. Lower values of the Brier loss $\in [0, 1]$ indicate better calibration.

$$\text{ECE} = \sum_{m=1}^M \frac{|B_m|}{N} |\text{acc}(B_m) - \text{conf}(B_m)| \quad (7.3)$$

$$\begin{aligned} \text{acc}(B_m) &= \frac{1}{|B_m|} \sum_{i \in B_m} y_i \\ \text{conf}(B_m) &= \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i \end{aligned}$$

$$\text{BCE Loss} = -\frac{1}{N} \sum_{i=1}^N [y_i \cdot \log(\hat{p}_i) + (1 - y_i) \cdot \log(1 - \hat{p}_i)] \quad (7.4)$$

$$\text{Brier Loss} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{p}_i)^2 \quad (7.5)$$

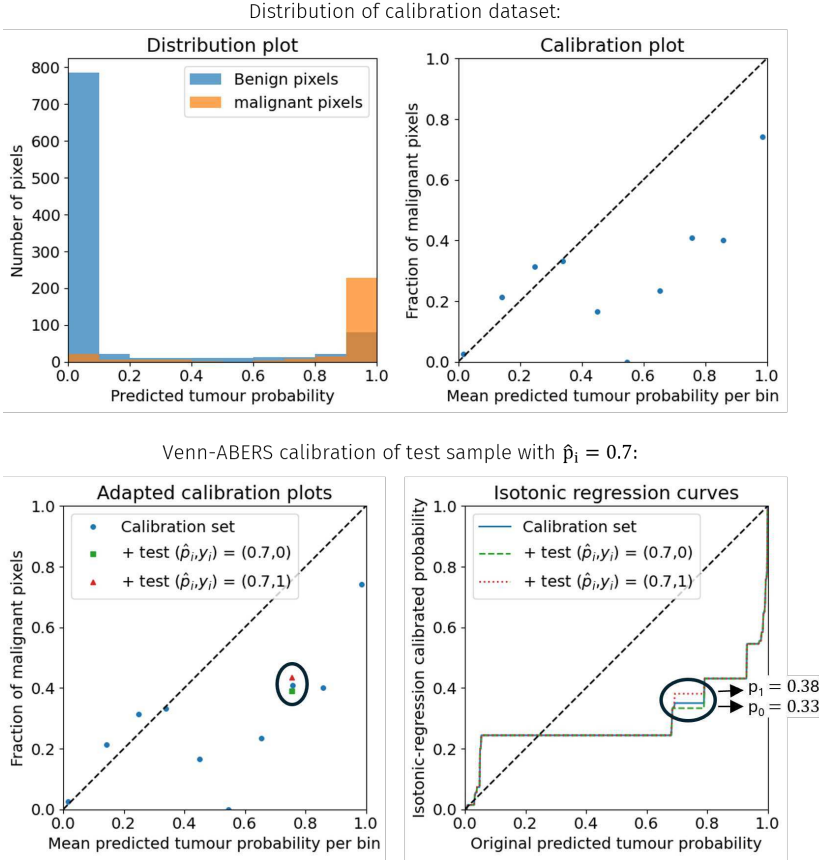


Figure 7.1. Illustration of inductive Venn-ABERS calibration for a single test pixel using one calibration fold. The top panels show the distribution (left) and calibration (right) of predicted tumour probabilities for all pixels in one calibration fold. The bottom panels illustrate the Venn-ABERS calibration for a single test pixel with predicted probability $\hat{p}_i = 0.7$. The test pixel is added to the calibration set twice, assuming labels $y_i = 0$ and $y_i = 1$. Isotonic regression is performed in both cases to obtain the lower and upper limit of the calibrated probability interval $[p_0, p_1]$ for pixel i . Black ovals indicate changes in the calibration plots (left) and isotonic regression curves (right) due to the inclusion of the test pixel. For visualisation purposes, only 1% of pixels from one validation fold are shown. This procedure needs to be performed for each calibration fold and each pixel.

7.2.3.2. Impact on segmentation and margin assessment performance

Furthermore, we evaluated whether tumour segmentation and margin assessment performance of the ResU-Nets were preserved after Venn-ABERS calibration. To evaluate tumour segmentation performance, predictions for the test set of 2D lumpectomy lamella images with ground-truth tumour segmentations were used. The predicted tumour probability maps were thresholded at probabilities in the range $[0, 1]$ to generate per-patient receiver operating characteristic (ROC) and precision-recall (PR) curves, reflecting pixel-level classification performance. The mean and standard deviation of the area under the curve of the receiver operating characteristic (ROC-AUC) and area under the curve of the precision-recall curve (PR-AUC) across patients were compared before and after calibration to evaluate the models' ability to distinguish between benign and malignant pixels. For evaluation of the margin assessment performance, tumour probability maps were first predicted for all micro-PET-CT images of lumpectomy specimens from the BrIMA dataset and subsequently calibrated using the Venn-ABERS method. As indicated in Section 6.2.2.1 of Chapter 6, there is a small overlap between the datasets with segmentation and margin status labels. Once again, it was ensured that tumour probability maps were never predicted using models trained or calibrated on data from the same patient. These tumour probability maps were then thresholded at probabilities in the range $[0, 1]$ to obtain tumour segmentations, from which margin status labels were derived following the procedure described in the previous chapter. Predicted and true margin status labels were compared to construct a single ROC- and PR-curve across all patients. The corresponding ROC-AUC and PR-AUC values were compared before and after calibration to evaluate the models' ability to discriminate between specimens with positive and negative margins.

7.2.3.3. Qualitative analysis of calibrated tumour probability maps

Following quantitative evaluation, we performed a visual comparison of tumour probability maps before and after Venn-ABERS calibration. Tumour probability maps were generated for several representative lumpectomy lamellas and several full specimens from the BrIMA dataset, to assess how calibration affects the representation of spatially-resolved prediction uncertainty. In addition, an alternative visualisation was created using the calibrated tumour probability maps. Tumour contours were drawn to delineate regions exceeding specific probability thresholds. For the selected cases, we examined whether these calibrated tumour probability maps and contours provided insight into tumour boundary locations relative to the specimen margins, and discussed the relation with the true margin status of the lumpectomy specimens.

7.3. Results

7.3.1. Quantitative calibration assessment

We applied inductive cross-Venn-ABERS calibration to the pixel-level tumour probabilities generated by the ResU-Nets using micro-PET-CT images as input. The calibration process was both fitted and evaluated on the dataset of NST lumpectomy lamella images with ground-truth tumour segmentations using the cross-validation scheme described above. This dataset contains approximately 1.3 million labelled tissue pixels, of which 14% are classified as malignant based on histopathology. Figure 7.2 presents both distribution and calibration plots for all predicted pixel-level tumour probabilities across the test set of 53 lamella images, shown before and after Venn-ABERS calibration. The initial calibration plot indicates that the uncalibrated ResU-Nets are overconfident, as the observed frequency of positive pixels falls below the mean predicted tumour probabilities. After applying Venn-ABERS calibration, the alignment improves substantially, although the resulting calibration plot does not perfectly match the diagonal. This enhancement in calibration is further supported quantitatively by reductions in all three calibration metrics (ECE, BCE loss, and Brier loss), as reported in Table 7.1.

Table 7.1. Calibration metrics for pixel-level tumour probabilities before and after Venn-ABERS calibration. Tumour probability maps were generated using the Residual U-Nets based on micro-PET-CT inputs. Reported values were calculated across approximately 1.3 million pixel-level tumour probabilities predicted for the test set of 53 lumpectomy lamella images from 19 patients with invasive carcinoma of no special type, with corresponding ground-truth tumour segmentations. Reductions in metric values are reported relative to the uncalibrated baseline values.

Metrics	Before calibration	After calibration	Relative reduction (%)
ECE	0.043	0.017	60.5
BCE Loss	0.209	0.172	17.7
Brier Loss	0.061	0.052	14.8

ECE = expected calibration error (calculated for 10 bins), BCE = binary cross-entropy

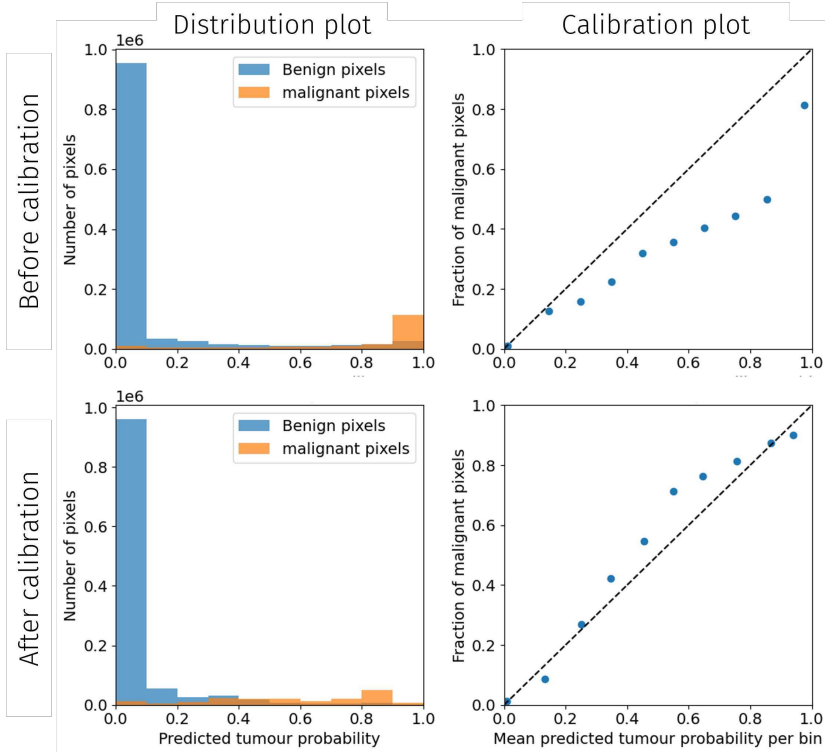


Figure 7.2. Distribution and calibration plots for pixel-level tumour probabilities before and after Venn-ABERS calibration. Tumour probability maps were generated using the Residual U-Nets based on micro-PET-CT inputs. The results are shown for approximately 1.3 million pixel-level tumour probabilities predicted for the test set of 53 lumpectomy lamella images from 19 patients with invasive carcinoma of no special type, with corresponding ground-truth tumour segmentations.

7.3.2. Impact on segmentation and margin assessment performance

Next, we evaluated the tumour segmentation and margin assessment performance of the calibrated ResU-Nets. Table 7.2 presents a comparison of ROC-AUC and PR-AUC values for both tasks before and after Venn-ABERS calibration. While tumour segmentation performance remains largely unchanged following calibration, margin assessment performance shows a slight decline with a lower PR-AUC value.

Table 7.2. Metrics expressing the tumour segmentation and margin assessment performance of the Residual U-Nets (ResU-Nets) before and after Venn-ABERS calibration. The ResU-Nets were applied to the micro-PET-CT inputs. For the segmentation performance, the average ($\pm$ standard deviation) of the ROC-AUC and PR-AUC across all 19 test patients are reported for the dataset of lumpectomy lamellas from invasive carcinomas of no special type (NST) with ground-truth tumour segmentations. For the margin assessment performance, the ROC-AUC and PR-AUC are reported for the BrIMA dataset of 84 NST lumpectomy specimens with ground-truth margin status.

Tumour segmentation		
Metrics	Before calibration	After calibration
ROC-AUC	0.96 $\pm$ 0.04	0.96 $\pm$ 0.03
PR-AUC	0.83 $\pm$ 0.14	0.82 $\pm$ 0.16
Margin assessment		
Metrics	Before calibration	After calibration
ROC-AUC	0.74	0.73
PR-AUC	0.39	0.30

ROC-AUC = area under the curve of the receiver operating characteristic, PR-AUC = area under the curve of the precision-recall curve

7.3.3. Qualitative analysis of calibrated tumour probability maps

Figure 7.3 compares the predicted tumour probability maps for four NST lamellas before and after Venn-ABERS calibration. While the uncalibrated tumour probability maps generally show overly confident sharp transitions between regions with high and low tumour probabilities, the calibrated maps provide a more nuanced representation. After calibration, the probability maps represent prediction uncertainty more reliably: regions where the model is confident about the tumour boundary are distinguished from broader transition zones that reflect higher uncertainty. This behaviour is also evident in the displayed tumour contours for different calibrated probability levels: closely spaced contours indicate confident tumour boundaries, whereas widely spaced contours highlight uncertainty.

Figures 7.4-7.7 present similar comparisons of tumour probability maps before and after Venn-ABERS calibration for four BrIMA specimens with NST tumours. The example images include specimens with a true positive (TP), true negative (TN), false positive (FP), and false negative (FN) margin status prediction based on a 50% threshold applied to the uncalibrated tumour probability maps as explained in the previous chapter. For the TP case in Figure 7.4, the calibrated tumour probability map shows a sharp transition near the specimen border, indicating high confidence in a positive margin. For the TN case in Figure 7.5, a broader transition zone is observed in the calibrated tumour probability map, but it remains distant from the specimen border, supporting a confident negative margin prediction. In contrast, the FP and FN cases in Figures 7.6 and 7.7, respectively, exhibit large transition zones extending to the specimen border, reflecting substantial uncertainty and suggesting that these predictions should be interpreted with caution.

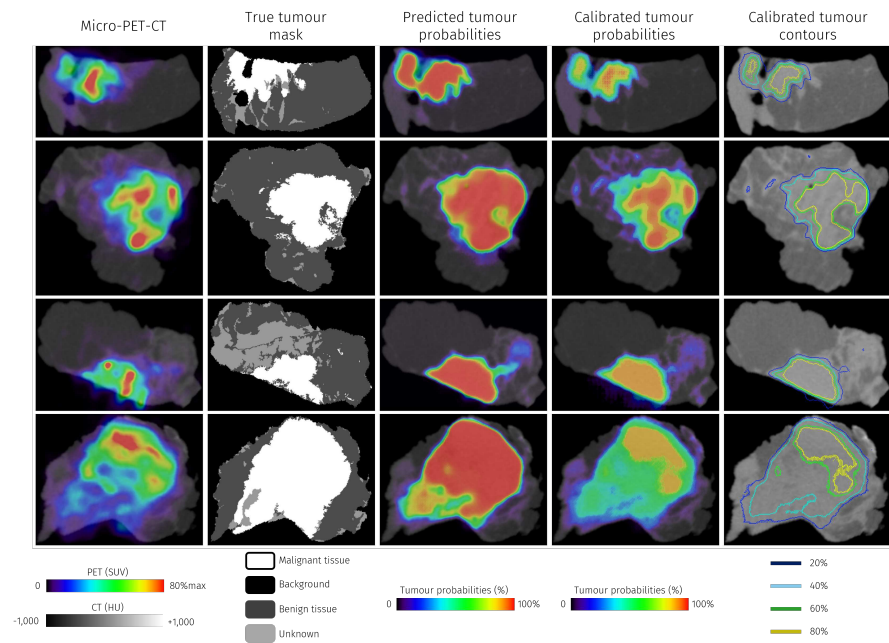


Figure 7.3. Visual comparison of tumour probability maps for lumpectomy lamellas before and after Venn-ABERS calibration. Four lamellas from four different patients with invasive carcinoma of no special type are presented (one per row). For each lamella, the micro-PET-CT image (expressed in standardised uptake values (SUVs) and Hounsfield units (HUs)) is displayed alongside the ground-truth tumour segmentation based on histopathology. Tumour probability maps generated by the Residual U-Nets from the micro-PET-CT inputs are shown before and after calibration. Tumour contour lines are shown, delineating regions with calibrated pixel-level tumour probabilities exceeding 20%, 40%, 60%, and 80%, respectively.

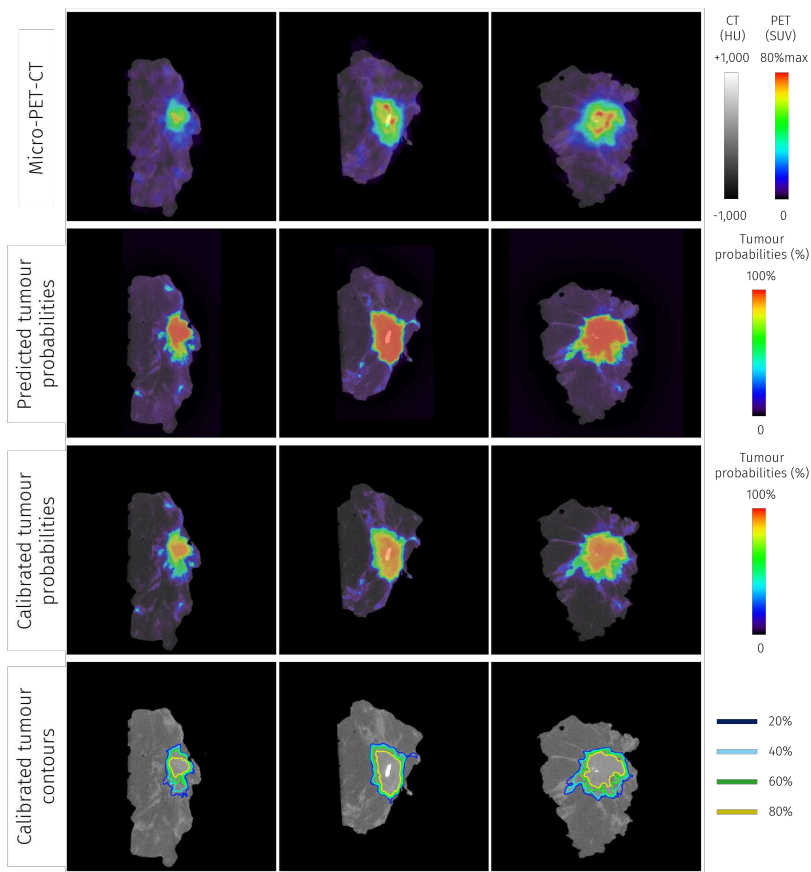


Figure 7.4. Visual comparison of tumour probability maps before and after Venn-ABERS calibration for a BrIMA lumpectomy specimen of invasive carcinoma of no special type with a positive margin according to histopathology. Three orthogonal slices are shown (one per column). For each slice, the micro-PET-CT input (expressed in standardised uptake values (SUVs) and Hounsfield units (HUs)) is displayed alongside tumour probability maps generated by the Residual U-Nets from the micro-PET-CT inputs, both before and after calibration. Tumour contour lines are shown, delineating the largest connected regions with calibrated pixel-level tumour probabilities exceeding 20%, 40%, 60%, and 80%, respectively. Applying a 50% threshold to the uncalibrated tumour probability map resulted in a true positive margin status prediction, following the procedure described in the previous chapter.

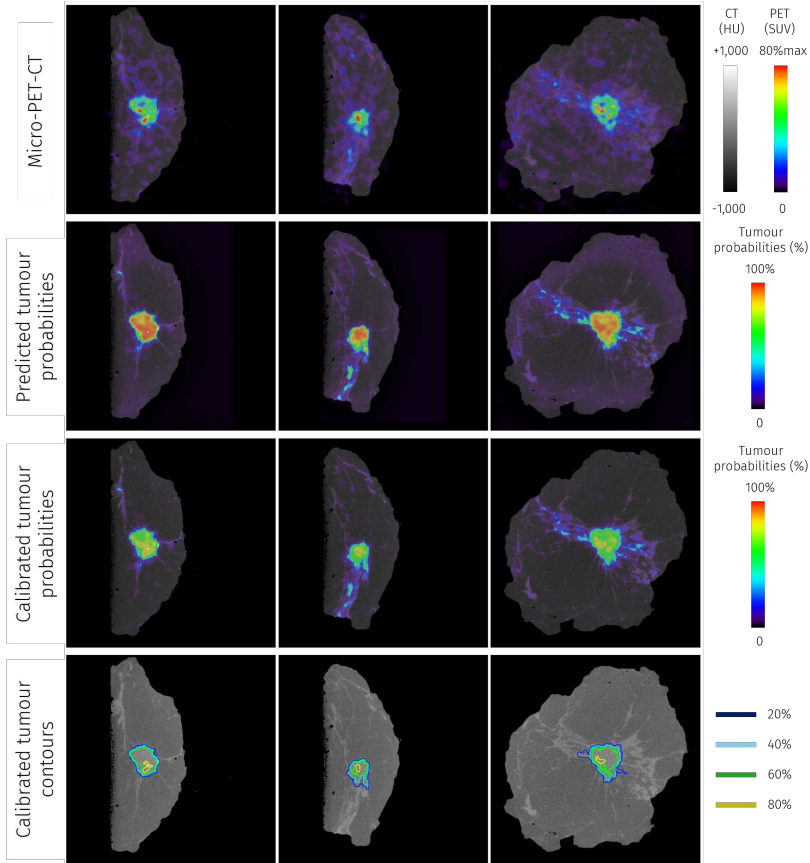


Figure 7.5. Visual comparison of tumour probability maps before and after Venn-ABERS calibration for a BrIMA lumpectomy specimen of invasive carcinoma of no special type with a negative margin according to histopathology. Three orthogonal slices are shown (one per column). For each slice, the micro-PET-CT input (expressed in standardised uptake values (SUVs) and Hounsfield units (HUs)) is displayed alongside tumour probability maps generated by the Residual U-Nets from the micro-PET-CT inputs, both before and after calibration. Tumour contour lines are shown, delineating the largest connected regions with calibrated pixel-level tumour probabilities exceeding 20%, 40%, 60%, and 80%, respectively. Applying a 50% threshold to the uncalibrated tumour probability map resulted in a true negative margin status prediction, following the procedure described in the previous chapter.

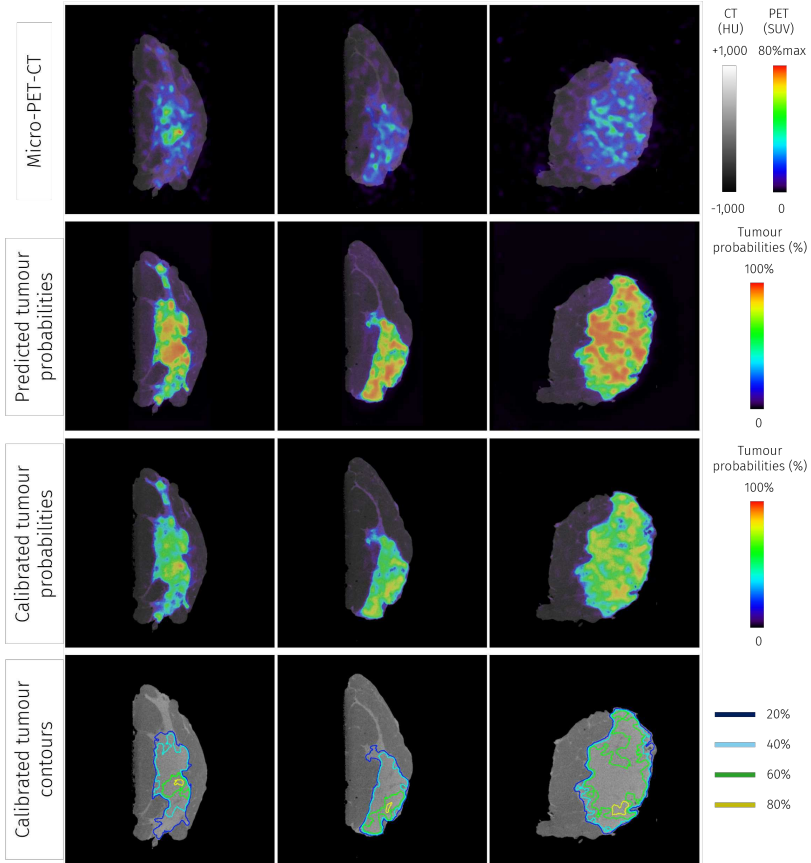


Figure 7.6. Visual comparison of tumour probability maps before and after Venn-ABERS calibration for a BrIMA lumpectomy specimen of invasive carcinoma of no special type with a negative margin according to histopathology. Three orthogonal slices are shown (one per column). For each slice, the micro-PET-CT input (expressed in standardised uptake values (SUVs) and Hounsfield units (HUs)) is displayed alongside tumour probability maps generated by the Residual U-Nets from the micro-PET-CT inputs, both before and after calibration. Tumour contour lines are shown, delineating the largest connected regions with calibrated pixel-level tumour probabilities exceeding 20%, 40%, 60%, and 80%, respectively. Applying a 50% threshold to the uncalibrated tumour probability map resulted in a false positive margin status prediction, following the procedure described in the previous chapter.

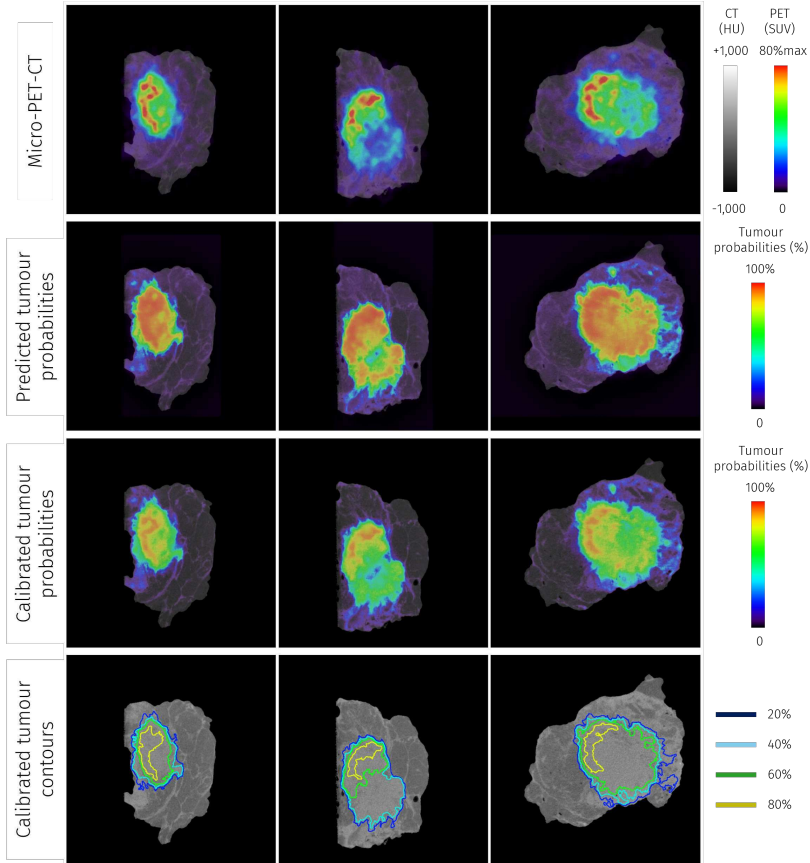


Figure 7.7. Visual comparison of tumour probability maps before and after Venn-ABERS calibration for a BrIMA lumpectomy specimen of invasive carcinoma of no special type with a positive margin according to histopathology. Three orthogonal slices are shown (one per column). For each slice, the micro-PET-CT input (expressed in standardised uptake values (SUVs) and Hounsfield units (HUs)) is displayed alongside tumour probability maps generated by the Residual U-Nets from the micro-PET-CT inputs, both before and after calibration. Tumour contour lines are shown, delineating the largest connected regions with calibrated pixel-level tumour probabilities exceeding 20%, 40%, 60%, and 80%, respectively. Applying a 50% threshold to the uncalibrated tumour probability map resulted in a false negative margin status prediction, following the procedure described in the previous chapter.

7.4. Discussion

The inductive cross-Venn-ABERS method was applied to calibrate tumour probability maps predicted from micro-PET-CT images of lumpectomy specimens using an ensemble of ResU-Nets. This resulted in a clear improvement in pixel-level probability calibration, with reductions of 60.5%, 17.7%, and 14.8% in ECE, BCE loss, and Brier loss, respectively. Calibration plots indicated that the Venn-ABERS method improved the alignment between the mean predicted tumour probabilities and the observed frequency of malignant pixels. Additionally, tumour segmentation performance of the DL models remained largely unchanged after calibration, while margin assessment performance showed only a slight decrease in PR-AUC. Furthermore, a qualitative comparison of the tumour probability maps before and after Venn-ABERS calibration, for lumpectomy lamellas and full specimens from the BrIMA dataset, showed that calibration produced more nuanced maps with broader transition zones indicating uncertain tumour boundaries. Overall, the tumour probability maps, together with their corresponding tumour contours, offered a more transparent and informative representation of model predictions than binary tumour segmentations or margin status labels, giving spatially-resolved insight into prediction uncertainty. Calibration then ensured that this uncertainty was reliably represented, thereby improving interpretability and fostering trust. Because the calibrated maps reflect the true likelihood of malignancy, they distinguish confidently positive, confidently negative, and genuinely ambiguous regions along the specimen margin, potentially allowing physicians to direct their attention to uncertain areas, thereby improving efficiency. In addition, reliable probability-based predictions might allow physicians to make risk-informed clinical decisions, tailored to the individual patient and clinical context.

In this work, we evaluated only a single calibration approach, namely Venn-ABERS calibration, although many alternative methods exist. Other post hoc techniques that could be investigated include Platt scaling and isotonic regression [119]. However, these non-conformal methods impose stricter requirements on data distribution, dataset size, and model flexibility, and they do not provide the same theoretical guarantees of valid calibration as Venn-ABERS [119, 183]. In addition, in many other DL applications, Venn-ABERS has outperformed Platt scaling and isotonic regression in terms of calibration performance [119, 184]. Furthermore, there are ad hoc calibration approaches that require retraining of the DL models. For example, Mehrtash et al. [181] demonstrated that using an ensemble of models instead of a single model can substantially improve calibration, an approach already adopted in our work. They also found that training with BCE loss, a proper scoring rule, can enhance calibration

compared to the Dice loss used in our study [181]. Future work could involve a systematic comparison of different calibration methods.

While the calibrated tumour probability maps presented in this work offer improved insight into prediction uncertainty regarding the location of the tumour boundary, they do not provide formal guarantees on tumour boundary location. Davenport et al. [186] and Mossina et al. [187] proposed patient-level conformal prediction methods, as opposed to our pixel-level approach, that yield spatial boundaries with statistical guarantees that the true tumour contour lies between them. This direction is promising for future research, however, these methods require substantially more data as each patient contributes only a single data point to the calibration set. According to Angelopoulos et al. [188] and Manokhin et al. [119], an ideal calibration set should contain around 500 data points and not fewer than 100 to ensure reliable calibration.

Although qualitative evaluation of the tumour probability maps showed promising results, further validation is required. In future work, we could develop a user-friendly interface that enables physicians to interact with calibrated tumour probability maps and tumour contours, allowing evaluation of the interpretability and usefulness of these maps for margin assessment. Hemon et al. [118] previously mentioned the importance of qualitative clinician feedback on the usefulness of probability maps. De Biase et al. [189] performed such a study in which they showed that DL-generated tumour probability maps based on PET-CT images of patients with head-and-neck tumours were perceived as intuitive and transparent by radiation oncologists to assist in gross tumour volume segmentation. As proposed by this group, interactive tools such as sliders could be incorporated to visualise tumour contours at different probability thresholds.

Our study contains several limitations. First, while Venn-ABERS improved calibration performance, the resulting calibration plot still deviated from the ideal diagonal. This discrepancy might be due to unmet assumptions of the Venn-ABERS method: (1) because of the limited dataset, validation folds were used for calibration instead of entirely independent calibration sets, and (2) pixel-level tumour probabilities within the same lumpectomy specimen are spatially correlated, violating the assumption of independent and identically distributed samples. Acquiring additional data with ground-truth tumour segmentation labels could enable the construction of truly independent calibration sets and potentially improve calibration performance. Second, calibration of the tumour probability maps for the full lumpectomy specimens was performed using lamella data. Although we believe that pixel-level probabilities from lamellas are exchangeable with voxel-level probabilities from full specimens, we could not quantitatively evaluate the calibration for full specimens. Such evaluation would require 3D micro-PET-CT data with corresponding ground-truth segmentation labels. Third, this work

only considered a single calibration approach. Future studies should compare multiple post hoc and ad hoc methods, both at the pixel and patient levels. Fourth, although representative examples were selected for qualitative assessment of calibrated tumour probability maps, the visualisations were limited to a small subset of the data and interpretation remains inherently subjective. Future work should assess the clinical interpretability and usefulness of calibrated tumour probability maps to aid physicians in margin assessment. Finally, calibration was performed only for NST lumpectomy specimens, as the number of data samples with ground-truth tumour segmentation labels for less common BCa subtypes, such as invasive lobular carcinoma (ILC) and ductal carcinoma in situ (DCIS), was too limited to support reliable calibration. Future work will focus on expanding the dataset to include these subtypes.

7.5. Conclusions

In this work, we aimed to enhance the trustworthiness of our DL models to aid physicians in the interpretation of micro-PET-CT images of lumpectomy specimens by moving from binary tumour segmentation to calibrated tumour probability maps. By applying inductive cross-Venn-ABERS calibration to calibrate pixel-level tumour probabilities, we substantially improved the alignment between predicted tumour probabilities and observed frequencies of malignant pixels. Tumour segmentation performance was largely preserved after calibration, and margin assessment performance was only slightly affected. Importantly, moving from binary segmentations to probability maps gave insight into spatially-resolved prediction uncertainty, and calibration made this uncertainty more reliable. The calibrated tumour probability maps and contours might therefore allow physicians to focus their attention on genuinely uncertain regions along the specimen border, potentially improving decision-making efficiency. Probabilistic and well-calibrated predictions might also allow physicians to tailor decisions to individual patients and specific clinical contexts.

This work takes a first step in closing the gap towards clinical deployment of a decision-support system to aid physicians in micro-PET-CT image interpretation. Future research should focus on expanding datasets to improve and validate calibration, exploring alternative calibration methods, and validating clinical usability and value of calibrated tumour probability maps to aid in margin assessment for lumpectomy specimens. Ultimately, improving model calibration and communicating prediction uncertainty towards physicians are key for enhancing trustworthiness and interpretability, and for enabling reliable deployment of the DL models in clinical practice.

Part III

General Conclusions and Future Perspectives

8 | Conclusions and future perspectives

8.1. Results summary

This dissertation explored the first steps towards a decision-support system to assist physicians in interpreting micro-PET-CT images of lumpectomy specimens for intraoperative margin assessment (IMA) during breast-conserving surgery (BCS). For this purpose, a deep learning (DL) model was developed that automatically segments breast tumours in micro-PET-CT images of lumpectomy specimens, thereby also enabling automated margin status prediction.

To develop such a DL model, we collected a dataset of micro-PET-CT images of breast cancer (BCa) specimens with ground-truth tumour segmentations. In Chapter 4, we proposed a method to co-register micro-PET-CT images of tissue lamellas with corresponding histopathology whole slide images (WSIs). The approach combined an image acquisition scheme that minimises tissue deformation with a mathematical co-registration algorithm that optimises image alignment. Compared with existing *in vivo* PET-histopathology alignment methods, the proposed workflow is less labour-intensive and time-consuming, requiring only fresh cutting of lamellas in addition to standard histopathology, and performing automatic co-registration using intensity-based similarity metrics. The method was applied to 53, 19, and 10 cassettes from 19 specimens of invasive carcinoma of no special type (NST), 6 of invasive lobular carcinoma (ILC), and 3 of ductal carcinoma in situ (DCIS), respectively, from patients preoperatively injected with [^{18}F]fluorodeoxyglucose ([^{18}F]FDG). Using microcalcifications as landmarks, the mean co-registration error between micro-PET and histopathology was 0.69 ± 0.23 mm. The co-registered 2D micro-PET-CT images and WSIs, annotated by a pathologist, provided insight into high-resolution [^{18}F]FDG uptake patterns. In general, higher mean standardised uptake values (SUVs) were found for (pre)malignant tissue (NST, ILC, DCIS) compared to healthy tissue (glandular, connective, and adipose). As expected, inflammation and skin also demonstrated high [^{18}F]FDG uptake.

The collected dataset of 2D [^{18}F]FDG micro-PET-CT images of lumpectomy lamellas with histopathology-derived ground-truth tumour segmentations was used in Chapter 5 to develop DL models for breast tumour segmentation. We used a stratified group nested five-fold cross-validation scheme to train and evaluate an ensemble of 2D Residual U-Nets (ResU-Nets) to delineate (pre)malignant tissue in micro-PET-CT images of BCa specimens. For the segmentation of NST tumours, the ResU-Nets reached a Dice similarity coefficient (DSC) of 0.73 ± 0.20 , and the area under the curve of the precision-recall curve (PR-AUC) was 0.83 ± 0.14 . The ResU-Nets

outperformed conventional PET intensity thresholding in terms of PR-AUC for the NST cases. The segmentation of ILC and DCIS tumours was found to be more challenging. Although the ResU-Nets performed better than intensity thresholding, due to the small datasets, we could not show statistically significant improvements for these BCa subtypes. Furthermore, we showed that combining micro-PET and micro-CT information is essential for accurate tumour delineation, with dual-modality input models outperforming single-modality input models.

In Chapter 6, we validated the ability of the ResU-Nets to accurately predict the margin status of BCa specimens. For this purpose, we used the 3D micro-PET-CT images of 41 and 152 early-stage BCa specimens (including NST, ILC, DCIS, and neoadjuvant-treated (NAT) tumours) collected in the BIMAP and BrIMA studies, respectively. The ensemble of ResU-Nets was used to segment (pre)malignant tissue in these previously unseen images, after which margin status was predicted based on whether the segmented tumour touched the specimen border. Comparing predicted and true histopathological margin status in all NST cases (BIMAP: $N=31$, BrIMA: $N=84$), the ResU-Nets reached an F1 score of 84% and 41% for the BIMAP and BrIMA datasets, respectively. For all BCa subtypes combined, F1 scores of 76% and 50% were obtained for the BIMAP and BrIMA datasets. Both for the NST tumours and for the combined group of all BCa subtypes in the BIMAP and BrIMA datasets, the ResU-Nets showed margin assessment performance statistically equivalent to intraoperative and postoperative physicians' assessments, while outperforming intensity thresholding. Although statistical significance could not be shown for some smaller ILC, DCIS, and NAT subsets, ResU-Nets generally showed higher specificity, precision, and F1 scores compared to intensity thresholding.

Safe clinical deployment of these DL models requires insight into prediction uncertainty. The tumour probability maps predicted by the ResU-Nets express this uncertainty at the pixel level, and Chapter 7 focused on calibrating them so that the predicted probabilities reflect the true likelihood of malignancy, thereby improving the trustworthiness and interpretability of the model outputs. Inductive cross-Venn-ABERS calibration was applied to these probability maps, focusing on the largest subgroup of NST BCa patients. Calibration was evaluated on the dataset of 53 2D micro-PET-CT images from lumpectomy lamellas of 19 patients with NST BCa, with true histopathology-derived tumour segmentations. Calibration substantially improved agreement between predicted pixel-level tumour probabilities and observed frequencies of malignant pixels. The expected calibration error (ECE), binary cross-entropy (BCE) loss, and Brier loss were reduced by 60.5%, 17.7%, and 14.8%, to 0.017, 0.172, and 0.052, respectively. Calibrated probability maps were also generated for the 84 3D micro-PET-CT images of NST specimens in the BrIMA dataset with known margin status labels. While segmentation performance remained stable after calibration, margin assessment showed a slight decrease in PR-AUC.

Qualitatively, the calibrated tumour probability maps provided reliable spatially-resolved uncertainty visualisations, enabling identification of uncertain tumour boundaries and ambiguous margin regions.

8.2. Future perspectives

We demonstrated a complete workflow to develop a DL-based decision-support system for micro-PET-CT specimen image interpretation, covering data collection, model development and evaluation, and communication of prediction uncertainty to physicians. However, this implementation represents only one possible approach, and improvements remain possible throughout the pipeline. In this section we discuss the most important directions for future work.

8.2.1. Data and ground truth development

8.2.1.1. Improved co-registration using high-contrast images

The limited soft tissue contrast in micro-CT images constrains co-registration accuracy. This could be improved by simultaneously capturing imaging modalities with higher contrast, such as MRI or contrast-enhanced CT, which can serve as intermediate references for more accurate alignment between histopathology and micro-PET-CT. Alternatively, such high-contrast imaging modalities could improve evaluation of co-registration accuracy as they allow easier identification of matching landmarks with histopathology, rather than relying on microcalcifications present in only a subset of images.

8.2.1.2. 3D micro-PET-CT data with ground-truth tumour segmentations

The proposed co-registration method generates 2D micro-PET-CT images with ground-truth breast tumour segmentations, whereas the ultimate objective is tumour segmentation in 3D images. The use of 2D data introduces several limitations. First, it constrains the accuracy of SUV analysis across breast tissue types, as the PET signal in 2D image slices depends on the surrounding 3D tissue configuration due to partial volume effects. Second, 3D labelled data could provide better insight into the micro-PET signature of ILC and DCIS tumours, which often exhibit infiltrative growth patterns that are difficult to characterise from a single 2D slice. Third, the dataset only enables training of 2D DL models, possibly limiting segmentation performance. In addition, using 2D images may limit the accuracy of segmentation performance evaluation, as the lack of

spatial relationships between slices prevents contour-based metrics, such as the Hausdorff distance and contour Dice, from properly evaluating the 3D structures. Finally, margin assessment performance may be affected by mismatches between the 2D lamella images used for training and the 3D full-specimen images used for validation, although this effect is expected to be minimal due to consistent imaging, reconstruction, and processing conditions. Similarly, although we believe that pixel-level tumour probabilities predicted for lamellas are exchangeable with voxel-level probabilities predicted for full specimens, calibration performance could not be evaluated for the full specimens. In future work, we plan to develop a method to co-register 3D micro-PET-CT images of entire specimens with a stack of the corresponding histopathology WSIs to obtain 3D ground-truth tumour segmentations. One possible approach is to first co-register the micro-PET-CT volume with a full stack of equally spaced histopathology WSIs, and then apply our slice matching technique from the 2D co-registration algorithm to further refine the alignment. This approach would eliminate the need to acquire micro-PET-CT images of tissue lamellas, thereby reducing the labour-intensity of the method. Moreover, it would broaden the applicability of the method, as it would no longer require fresh cutting of specimens, thereby causing less disruption to the standard histopathological workflow.

8.2.1.3. Improved insight into radiotracer uptake in breast tumours

The dataset of co-registered micro-PET-CT images and histopathology WSIs contains substantial unexplored information. We observed a wide range of SUVs across patients with the same histological tumour type, and tumours frequently exhibited heterogeneous [^{18}F]FDG uptake. Beyond histological BCa type, several other factors may influence tumour SUVs, including molecular subtype, tumour grade, cell density, microvessel density, and the presence of tumour infiltrating immune cells. Future work may investigate how these tumour characteristics relate to radiotracer uptake, using these insights to improve interpretation of micro-PET-CT images. Moreover, the co-registered micro-PET-CT images and WSIs provide a unique opportunity to study the tumour microenvironment and the biological drivers of metabolic heterogeneity, which may ultimately contribute to more personalised patient care.

8.2.2. Improving model performance

8.2.2.1. Data-efficient and collaborative learning strategies

To further improve, validate, and calibrate the proposed DL models, additional micro-PET-CT images with ground-truth tumour segmentations

and margin status labels should be collected. This is particularly important for ILC and DCIS tumours, where achieving clear margins remains challenging in clinical practice. However, since collecting such data is time-consuming and labour-intensive, alternative data-efficient strategies should be explored to improve model performance. One option is synthetic data generation using generative adversarial networks (GANs) or diffusion models, as discussed in Section 3.3.5.1 of Chapter 3. Another approach to obtain artificial data is to create a digital twin of the micro-PET-CT scanner and use Monte Carlo simulations to generate realistic images of modelled lumpectomy specimens. In addition, data-efficient learning strategies may further improve performance. Transfer learning from existing pre-trained models, such as ImageNet or MedSAM, could already boost model performance without requiring additional labelled data. Self-supervised learning, as introduced in Section 3.3.5.1 of Chapter 3, can also be used to exploit micro-PET-CT images without tumour annotations to learn meaningful feature representations before fine-tuning the models using labelled data. Finally, as previously discussed in Section 3.3.5.1 of Chapter 3, active learning could improve labelling efficiency by iteratively selecting the most informative specimen images for annotation, while federated learning could expand the effective training set by enabling multicentre collaboration without the need to exchange patient data.

8.2.2.2. Alternative model architectures

While currently a standard ResU-Net architecture was used for automatic breast tumour segmentation, alternative architectures could be explored. Vision transformers, introduced in Section 3.3.5.2 of Chapter 3, have recently gained popularity due to their strong global context modelling capabilities. However, they typically require large training datasets and are computationally demanding. Applying them to the current problem would therefore likely require selecting a suitable pre-trained model and fine-tuning it on the limited available labelled dataset. In addition, the current pipeline trains a tumour segmentation model, from which the output is subsequently used to determine the margin status for lumpectomy specimens. An alternative approach could be a multi-task model that simultaneously performs tumour segmentation and margin status prediction, as described in Section 3.3.5.2 of Chapter 3. Joint learning of these tasks might encourage the model to develop a more robust shared latent representation of the input image. However, such a model would require careful regularisation to avoid inconsistencies between the predicted segmentation and margin status classification. Moreover, this approach reduces pipeline interpretability, as the mechanism underlying the margin status prediction becomes less transparent and the model may rely on incorrect imaging features for classification.

8.2.2.3. Improved optimisation and supervision

Exploring alternative optimisation strategies might further improve the DL model's performance. While a standard Dice loss was used in this work, deep supervision (introduced in Section 3.3.5.3 of Chapter 3) could encourage the model to learn more meaningful intermediate data representations. In addition, learnable loss functions based on GANs or auto-encoder frameworks (introduced in Section 3.3.5.3 of Chapter 3), might promote structurally and anatomically plausible segmentations. Particularly for less common BCa types, such as ILC and DCIS, which often present with smaller tumour volumes and infiltrative growth patterns, more advanced loss functions might be required to capture higher-order structural similarities beyond pixel-wise accuracy.

8.2.3. Trustworthy and efficient deployment in clinical practice

Before this DL-based software could be integrated into the micro-PET-CT specimen scanner and deployed in clinical practice, several additional steps are required. First, further evaluation of the model's clinical outputs may be valuable. While currently we only evaluated the ability of the DL models to accurately predict the specimen-level margin status, it might also be important to assess whether the models can correctly identify the orientation of positive margins, which is clinically relevant for guiding cavity shaving. However, such evaluation is challenging because positive margin orientation can be subjective and as discussed in Section 2.3.4 of Chapter 2, discrepancies between surgeons' and pathologists' assessments are common due to specimen deformation or disorientation. Furthermore, model outputs should be presented in an intuitive and interpretable way that supports, rather than replaces, clinical decision-making. Transparency in prediction uncertainty is essential to identify potential errors and enhance trustworthiness. While we presented one possible approach for calibrating the tumour probability maps to improve the reliability of the spatially-resolved uncertainty they express, future work should compare alternative post hoc and ad hoc calibration methods. Conformal prediction at the patient level is a promising approach for generating valid tumour segmentation boundaries, although it requires large datasets for reliable calibration. Additionally, physician feedback is needed to assess interpretability and usefulness of these calibrated tumour probability maps, and determine which visualisations best support faster and more accurate decisions. Aspects like uncertainty quantification, model trustworthiness, and model-surgeon interaction are also emphasised in the artificial intelligence (AI) Act and are important considerations for regulatory approval. Finally, a more efficient and lightweight implementation of the DL model will be required to enable fast intraoperative breast tumour segmentation in 3D micro-PET-CT images

without extensive computational resources. One approach could be to investigate whether comparable segmentation performance can be achieved when running inference on downsampled micro-PET-CT images. Another option is to apply knowledge distillation, as introduced in Section 3.3.5.4 of Chapter 3, to compress the knowledge of the trained ensemble of ResU-Nets into a single lightweight model.

8.2.4. Regulatory considerations

Clinical adoption of the proposed DL-based decision-support software would require Conformité Européenne (European Conformity) (CE) marking in Europe. In its current form, the AURA[®]10 by XEOS performs micro-PET-CT imaging of resected specimens and is regulated under the In Vitro Diagnostic Regulation (IVDR) as a low-risk Class A device. Extending its intended purpose to margin assessment, with AI software guiding intraoperative decisions, would substantially raise the associated risk. As the software would analyse data derived from resected specimens, it would remain within the scope of the IVDR and qualify as Software as a Medical Device (SaMD). Under IVDR Annex VIII, software intended for the screening, diagnosis, or staging of cancer falls under classification Rule 3(h) and is assigned to Class C. Unlike the current Class A device, a Class C device requires notified-body approval before market entry, followed by annual surveillance assessments and approval of any significant software changes, together with a performance evaluation covering scientific validity, analytical performance, and clinical performance. Alongside the IVDR, two further frameworks would also apply: the General Data Protection Regulation (GDPR), which governs the protection of personal health data, and the AI Act, which emphasises the need for continuous human oversight and the use of unbiased, high-quality data for model training and testing, among other requirements. This human-oversight requirement aligns with the intended role of the software, which supports physicians to make intraoperative decisions and does not replace definitive margin confirmation by postoperative histopathology. It should be noted that the above classifications and requirements reflect preliminary assumptions rather than definitive conclusions, and would need further investigation and confirmation with regulatory experts, particularly as AI legislation continues to evolve.

8.2.5. Extension towards other cancer types and radiotracers

Although this work focused on [¹⁸F]FDG micro-PET-CT images of BCa specimens, the same strategy could be used to develop DL models to support micro-PET-CT interpretation for other cancer types and radiotracers. While images for different tumour types and radiotracers share

visual similarities, differences in intensity, contrast, radiotracer uptake patterns, and specimen morphology may still require model retraining or fine-tuning. Ultimately, a foundation model could be trained on a diverse set of micro-PET-CT images from different cancerous resection specimens with different radiotracers, requiring only limited labelled datasets to fine-tune the model for new applications. Nevertheless, labelled data will still be required to refine and validate models for other specimen types and radiotracers. Although feasibility of the proposed co-registration method was demonstrated in highly deformable breast tissue, other specimens may introduce different challenges. For instance, prostate specimens often show more homogeneous micro-CT images which may lack internal anatomical structures to properly guide alignment. In head-and-neck specimens, the presence of bone prevents fresh cutting of lamellas, further motivating the need for 3D co-registration algorithms. Another limitation of the current implementation is that the BCa subtype must be known beforehand, typically through biopsy, in order to select the appropriate model and obtain optimal segmentation results. Future work could aim to automate this step. In addition, the current models cannot distinguish between benign and malignant specimens. Due to relative scaling of micro-PET intensities, noise in benign specimens may be amplified, which can result in the DL models incorrectly detecting a tumour.

8.3. Conclusions

In this PhD dissertation, we have taken important steps towards the development of a decision-support system to aid physicians in interpreting micro-PET-CT images of lumpectomy specimens. First, we developed a novel method to accurately and automatically co-register micro-PET-CT specimen images with corresponding histopathology WSIs to collect a unique dataset of 2D [^{18}F]FDG micro-PET-CT images of breast tissue aligned with histopathology. We showed that, even for highly deformable breast tissue, co-registration can be successfully performed with an accuracy comparable to the spatial resolution of micro-PET. The collected dataset demonstrated that (pre)malignant breast tissue in general shows higher [^{18}F]FDG uptake compared to healthy tissue, also including the ILC and DCIS subtypes that are more difficult to detect in *in vivo* PET. We believe that the proposed technique can also be applied to other types of tumours and (new) radiotracers to provide a better understanding of the *in vivo* radiotracer distribution in human tissues on a submillimetre scale. These insights might eventually improve clinical practice and patient care. Afterwards, a DL-based segmentation model was developed using the collected dataset of micro-PET-CT images with histopathology-derived ground-truth breast tumour segmentations. The DL model outperformed conventional intensity thresholding for

segmentation of NST tumours, and showed similar trends for ILC and DCIS. Furthermore, it was shown that combining micro-PET and micro-CT information is essential for accurate tumour delineation. A unique aspect of our study is the use of histopathology instead of manual expert annotations as the segmentation gold standard, providing a more clinically relevant measure of tumour extent. Next, we validated the ability of the DL models to correctly predict the margin status from 3D micro-PET-CT images of lumpectomy specimens. The DL models achieved margin assessment performance statistically equivalent to intraoperative and postoperative physicians' evaluations, while consistently outperforming intensity thresholding for NST tumours and for the combined group of BCa subtypes. Notably, this equivalence was demonstrated in a large multicentre dataset with diverse BCa subtypes, reflecting clinical reality. Finally, we improved the trustworthiness of our DL models by replacing binary tumour segmentations with calibrated tumour probability maps for NST tumours. Calibration improved the alignment between predicted pixel-level probabilities and the true frequency of malignant pixels, while largely preserving segmentation and margin assessment performance. The calibrated maps provided a transparent visualisation of spatially-resolved prediction uncertainty, helping physicians identify ambiguous margin regions and potentially make more efficient, risk-informed decisions.

These findings highlight the potential of the proposed DL models to predict both tumour segmentations and margin status labels, as well as communicate spatially-resolved prediction uncertainty. Together, these capabilities may improve consistency and reduce decision-making time when interpreting micro-PET-CT images of lumpectomy specimens. Although the results look promising, more extensive validation of the DL models is required before they can be adopted in clinical practice. Larger and more diverse datasets, particularly including rarer BCa types such as ILC and DCIS, are needed to further improve and evaluate the models. In addition, future work should focus on improving model efficiency for fast intraoperative decision-making, as well as on further enhancing model trustworthiness to support effective physician-model collaboration. The clinical usability and value of the proposed DL models should still be investigated. While underscoring the need for further (clinical) validation, overall, the results highlight the DL model's potential as a first step towards a decision-support system to enhance IMA during BCS based on micro-PET-CT, facilitating the adoption of this technique in clinical practice.

Part IV

Appendices

A | Co-registration parameters

This appendix includes the full parameter files used for rigid, affine, and B-spline co-registration. For all registration parameters that are not explicitly mentioned, the default values from the Python SimpleElastix library are used [145]. Metrics are indicated in *italic* as they depend on the co-registration step as indicated in Table 4.1 in Chapter 4.

Rigid transformation:

```
(AutomaticParameterEstimation "true")
(AutomaticScalesEstimation "true")
(CheckNumberOfSamples "true")
(DefaultPixelValue -1000)
(FinalBSplineInterpolationOrder 1)
(FixedImagePyramid "FixedSmoothingImagePyramid")
(ImagePyramidSchedule 8 4 2 1)
(ImageSampler "RandomCoordinate")
(Interpolator "LinearInterpolator")
(MaximumNumberOfIterations 500)
(MaximumNumberOfSamplingAttempts 8)
(Metric "AdvancedNormalizedCorrelation")
(Metric "NormalizedMutualInformation")
(MovingImagePyramid "MovingSmoothingImagePyramid")
(NewSamplesEveryIteration "true")
(NumberOfResolutions 4)
(NumberOfSamplesForExactGradient 4096)
(NumberOfSpatialSamples 2048)
(Optimiser "AdaptiveStochasticGradientDescent")
(Registration "MultiResolutionRegistration")
(ResampleInterpolator "FinalBSplineInterpolator")
(Resampler "DefaultResampler")
(Transform "EulerTransform")
(UseDirectionCosines "true")
```

Affine transformation:

```
(AutomaticParameterEstimation "true")
(AutomaticScalesEstimation "true")
(CheckNumberOfSamples "true")
(DefaultPixelValue -1000)
(FinalBSplineInterpolationOrder 1)
(FixedImagePyramid "FixedSmoothingImagePyramid")
(ImagePyramidSchedule 8 4 2 1)
(ImageSampler "RandomCoordinate")
```

(Interpolator "LinearInterpolator")
(MaximumNumberOfIterations 500)
(MaximumNumberOfSamplingAttempts 8)
(Metric "AdvancedNormalizedCorrelation")
(Metric "NormalizedMutualInformation")
(MovingImagePyramid "MovingSmoothingImagePyramid")
(NewSamplesEveryIteration "true")
(NumberOfResolutions 4)
(NumberOfSamplesForExactGradient 4096)
(NumberOfSpatialSamples 2048)
(Optimiser "AdaptiveStochasticGradientDescent")
(Registration "MultiResolutionRegistration")
(ResampleInterpolator "FinalBSplineInterpolator")
(Resampler "DefaultResampler")
(Transform "AffineTransform")
(UseDirectionCosines "true")

B-spline transformation:

(AutomaticParameterEstimation "true")
(CheckNumberOfSamples "true")
(DefaultPixelValue -1000)
(FinalBSplineInterpolationOrder 1)
(FinalGridSpacingInPhysicalUnits 40)
(FixedImagePyramid "FixedSmoothingImagePyramid")
(GridSpacingSchedule 6 4 2 1)
(ImagePyramidSchedule 8 4 2 1)
(ImageSampler "RandomCoordinate")
(Interpolator "LinearInterpolator")
(MaximumNumberOfIterations 500)
(MaximumNumberOfSamplingAttempts 8)
(Metric "AdvancedNormalizedCorrelation" "TransformBendingEnergyPenalty")
(Metric "NormalizedMutualInformation" "TransformBendingEnergyPenalty")
(Metric0Weight 0.5)
(Metric1Weight 0.5)
(MovingImagePyramid "MovingSmoothingImagePyramid")
(NewSamplesEveryIteration "true")
(NumberOfResolutions 4)
(NumberOfSamplesForExactGradient 4096)
(NumberOfSpatialSamples 2048)
(Optimiser "AdaptiveStochasticGradientDescent")
(Registration "MultiMetricMultiResolutionRegistration")
(ResampleInterpolator "FinalBSplineInterpolator")
(Resampler "DefaultResampler")
(Transform "RecursiveBSplineTransform")

B | Impact of fixated micro-CT guidance on co-registration accuracy

This appendix shows the results of an analysis performed to investigate the added value of using the high-resolution (micro-)CT image of the fixated lamella as guidance during the co-registration procedure (step I). For this purpose, we studied the impact of omitting step I of the co-registration procedure on the final result for four lamellas. The resulting co-registered micro-CT images with and without step I were compared qualitatively and quantitatively. For this analysis, we selected the four lamellas containing corresponding microcalcifications in the final co-registered micro-CT images and whole slide images (WSIs), allowing to calculate the spatial co-registration error.

Figure B.1 shows the final micro-CT images of these four fresh lamellas co-registered with the WSIs with and without guidance of the micro-CT of the fixated lamellas. Only small differences can be detected in these images. Table B.1 shows quantitative metrics expressing the final alignment between the micro-CT images and WSIs after co-registration with and without guidance of the fixated micro-CT. Only a minor decrease is seen in the Dice similarity coefficient (DSC) evaluating the overlap in tissue composition segmentations between the co-registered micro-CT images and WSIs. When calculating the distances between corresponding microcalcifications in the co-registered micro-CT images and WSIs, we found a slight decrease in the mean spatial co-registration error: 0.49 ± 0.33 mm versus 0.47 ± 0.25 mm, with and without guidance of the fixated micro-CT, respectively.

Considering these qualitative and quantitative results, we conclude that using the micro-CT of the fixated lamella to guide the co-registration procedure only has a minor influence on the final result. To make the co-registration process less labour-intensive, the acquisition of the micro-CT images of the fixated lamellas can be safely omitted.

Table B.1. Quantitative metrics expressing final alignment between the micro-CT images and histopathology whole slide images after co-registration with and without guidance of the micro-CT of the fixated lamellas. The mean $\pm$ standard deviation of all quantitative metrics across four selected lamellas are shown after final co-registration.

	NMI	DSC _{shape}	DSC _{comp}
With Fixated CT	0.28 $\pm$ 0.02	0.95 $\pm$ 0.01	0.79 $\pm$ 0.08
Without Fixated CT	0.28 $\pm$ 0.02	0.95 $\pm$ 0.01	0.77 $\pm$ 0.08

NMI = normalised mutual information, DSC = Dice similarity coefficient

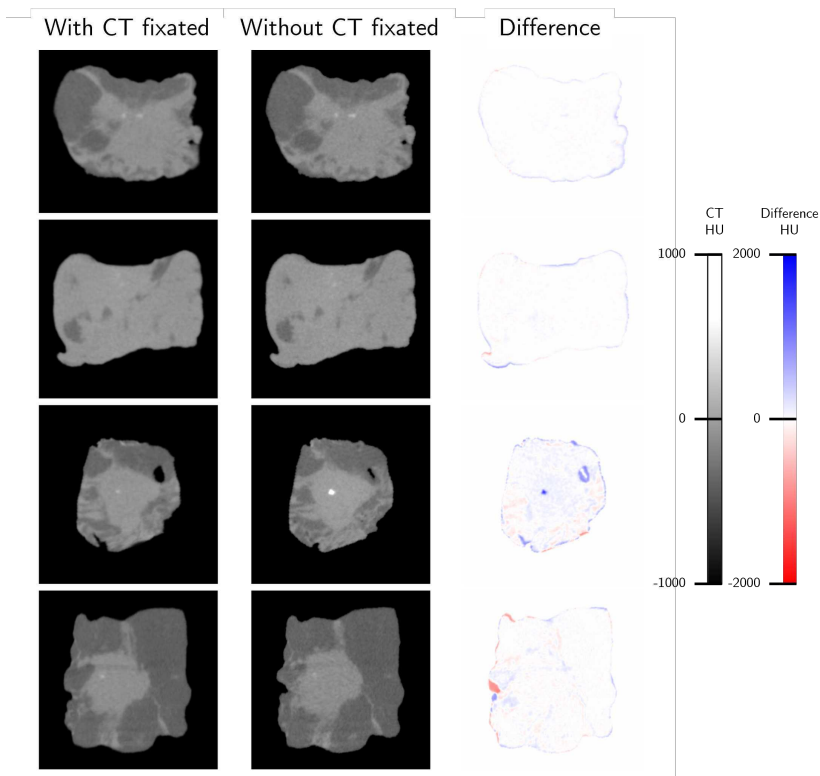


Figure B.1. Difference between micro-CT images of fresh lamellas co-registered with the histopathology whole slide images with and without guidance of the micro-CT of the fixated lamellas. Results are shown for four lamellas of different patients.

C | ResU-Net hyperparameters and learning curves

This appendix contains details on the chosen model hyperparameters and learning curves for the 2D Residual U-Nets (ResU-Nets) trained to segment (pre)malignant breast tissue in micro-PET-CT images of lumpectomy specimens. Table C.1 shows the chosen hyperparameters of the different ResU-Nets, optimised through grid search. Figures C.1 and C.2 show the evolution of the train and validation loss across epochs for training of the ResU-Nets on the invasive carcinomas of no special type (NST) and on the combined cases of invasive lobular carcinoma (ILC) and ductal carcinoma in situ (DCIS), respectively.

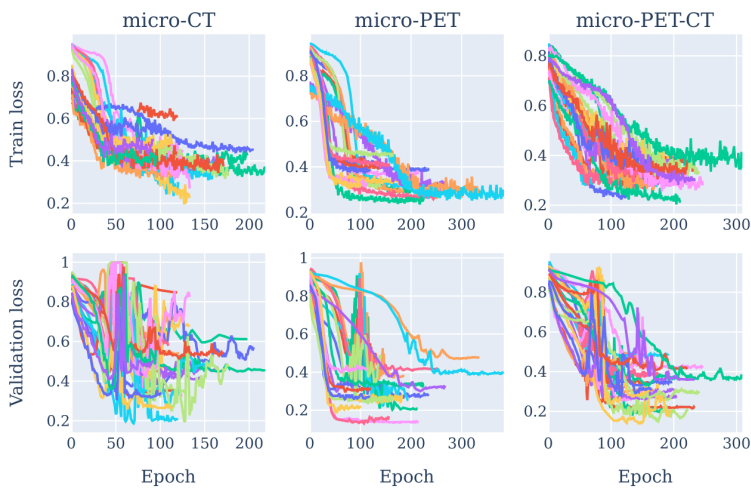


Figure C.1. Evolution of training and validation Dice loss across epochs during the training of 75 Residual U-Nets (ResU-Nets) on data from cases with invasive carcinoma of no special type (NST). Training curves are grouped by model input (micro-CT, micro-PET, and micro-PET-CT). Each plot displays 25 training curves, corresponding to 25 models trained in a five-fold nested cross-validation scheme.

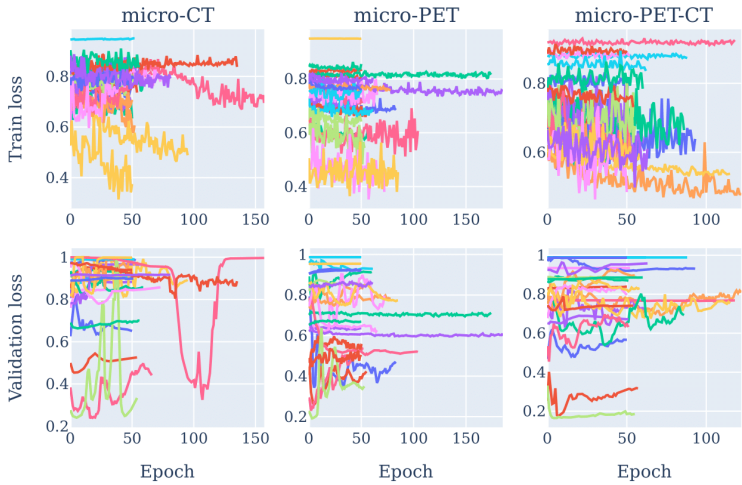


Figure C.2. Evolution of training and validation Dice loss across epochs during the transfer learning of 75 Residual U-Nets (ResU-Nets) on data from cases with invasive lobular carcinoma (ILC) and ductal carcinoma in situ (DCIS). Training curves are grouped by model input (micro-CT, micro-PET, and micro-PET-CT). Each plot displays 25 training curves, corresponding to 25 models trained in a five-fold nested cross-validation scheme.

Table C.1. Overview of hyperparameters of the Residual U-Nets (ResU-Nets) obtained through grid search. Hyperparameters were defined per model input and per test fold. The same hyperparameters were used when training the models using only the invasive carcinomas of no special type (NST), and when performing transfer learning for the cases of invasive lobular carcinoma (ILC) and ductal carcinoma in situ (DCIS). The batch size indicates the number of distinct images that is part of one batch, and should be multiplied with the number of patches per image to obtain the final batch size.

Input:	micro-CT					micro-PET					micro-PET-CT				
Test fold:	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
# layers	5	5	5	6	2	4	4	4	5	4	4	4	4	4	5
# channels	16	16	8	16	32	32	8	8	16	32	8	32	16	16	8
kernel size	5	5	5	5	5	5	3	5	7	5	3	5	3	3	5
normalisation type	B	B	B	B	L	B	B	B	B	B	B	B	B	B	B
dropout ratio	0.05	0.05	0.05	0.05	0.0	0.05	0.0	0.05	0.05	0.1	0.05	0.0	0.05	0.05	0.05
learning rate	1e-3	1e-2	1e-2	1e-2	1e-3	1e-3	1e-3	1e-3	1e-3	1e-3	1e-2	1e-2	1e-2	1e-2	1e-2
batch size	2	8	8	6	6	2	10	2	6	4	6	10	6	6	4
patch size	160	320	160	320	60	240	80	240	320	240	160	80	240	160	160
# patches per image	2	4	4	4	8	4	6	4	4	4	4	6	2	2	4
positive patch bias	1	2	1	2	2	1	2	1	1	1	2	1	1	2	2
L2 penalty	0.1	0.1	0.2	0.0	0.2	0.0	0.1	0.0	0.1	0.1	0.1	0.2	0.1	0.1	0.1

layers = number of down- and upsampling layers in the network

channels = number of channels in the first layer

kernel size = side length of the square kernel in pixels

normalisation type: B = batch, L = layer

patch size = side length of square image patches in pixels

patches per image = number of random patches extracted per image every epoch

positive patch bias = bias towards selecting a patch with a positive pixel at the centre:

1 = patches are selected at random, 2 = the chance of selecting a patch with a positive centre pixel is twice as high

D | Tumour segmentation performance for micro-PET-CT images of ILC and DCIS lumpectomy specimens

This appendix contains the full test results for the performance of the Residual U-Nets (ResU-Nets) and intensity thresholding models when segmenting tumours of invasive lobular carcinoma (ILC) or ductal carcinoma in situ (DCIS) in micro-PET-CT images. Visualisations of predicted tumour segmentations for a representative ILC and DCIS case can be found in Figures D.1 and D.2, respectively. The mean metric values (Dice similarity coefficient (DSC), area under the curve of the precision-recall curve (PR-AUC), 95th percentile of the Hausdorff distance (95HD), and contour Dice with a 1 mm tolerance (1mmCD)) across the test patients with ILC and DCIS tumours are shown in Table D.1. The box plots in Figures D.3 and D.4 show the distribution of the metric values in more detail for the ILC and DCIS tumours, respectively. In addition, Table D.1 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. Furthermore, Table D.2 shows the p-values resulting from the Wilcoxon signed-rank test 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.

Given the limited number of data samples for these breast cancer (BCa) types, the results should be interpreted with caution. Both for ILC and DCIS tumours, best performance was found when combining micro-PET with micro-CT to predict tumour segmentations. However, differences in performance were not statistically significant. While in general, ResU-Net performed better than intensity thresholding, once more no statistically significant differences were found. Tumour segmentation is particularly challenging for the DCIS cases, where we found a low performance in terms of DSC and PR-AUC.

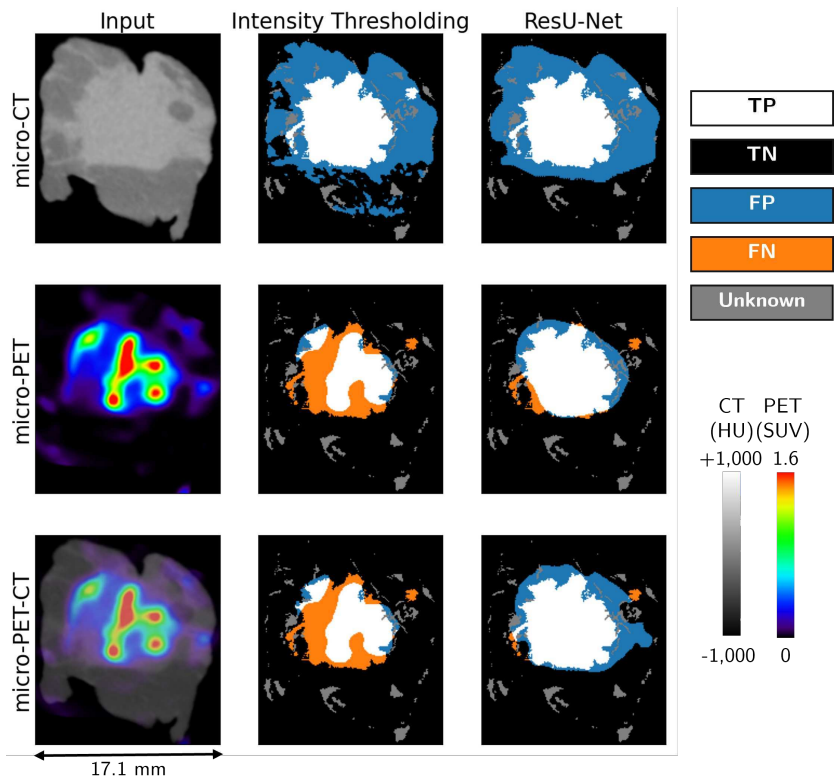


Figure D.1. Example of the tumour segmentations predicted with intensity thresholding and Residual U-Net (ResU-Net) for an invasive lobular carcinoma. Predictions are shown for three different inputs: micro-CT, micro-PET, and micro-PET-CT. The micro-PET and micro-CT images are expressed in standardised 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.

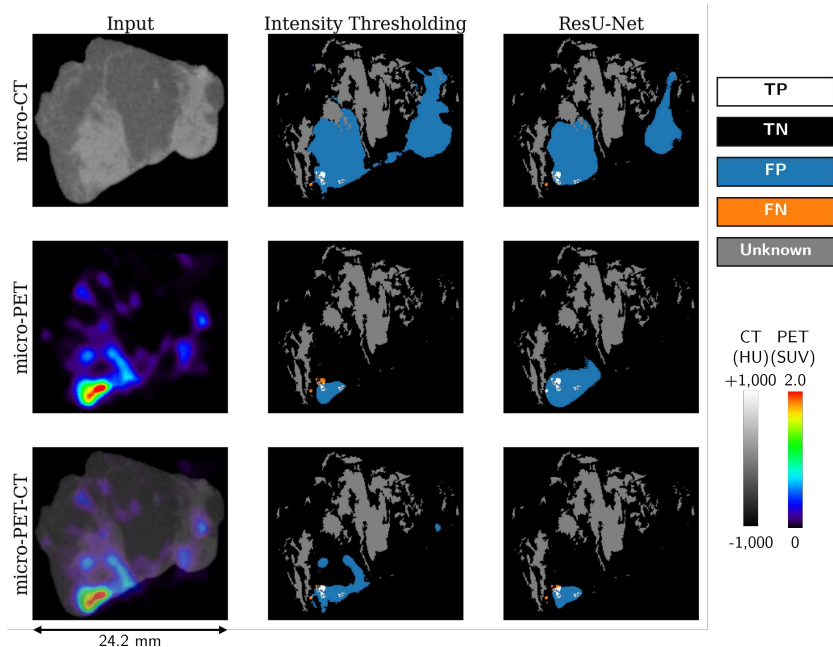


Figure D.2. Example of the tumour segmentations predicted with intensity thresholding and Residual U-Net (ResU-Net) for a ductal carcinoma in situ. Predictions are shown for three different inputs: micro-CT, micro-PET, and micro-PET-CT. The micro-PET and micro-CT images are expressed in standardised 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.

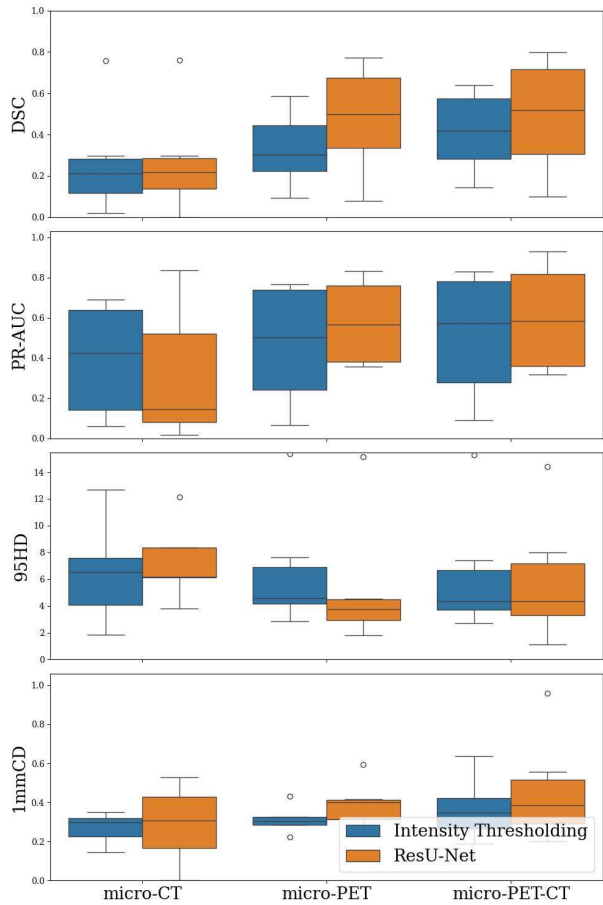


Figure D.3. Box plots showing metric distributions assessing agreement between true and predicted tumour segmentations for test patients with invasive lobular carcinoma. The following metrics are shown: Dice similarity coefficient (DSC), area under the curve of the precision-recall curve (PR-AUC), 95th percentile of the Hausdorff distance (95HD) in mm, and 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)).

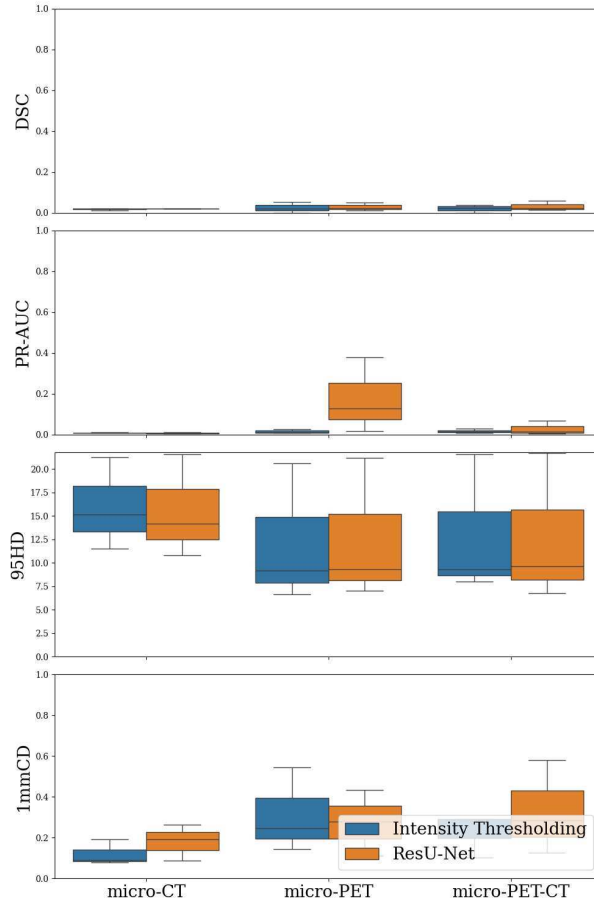


Figure D.4. Box plots showing metric distributions assessing agreement between true and predicted tumour segmentations for test patients with ductal carcinoma in situ. The following metrics are shown: Dice similarity coefficient (DSC), area under the curve of the precision-recall curve (PR-AUC), 95th percentile of the Hausdorff distance (95HD) in mm, and 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)).

Table D.1. Metric values to evaluate the agreement between true and predicted tumour segmentations for test patients with invasive lobular carcinoma (ILC) and ductal carcinoma in situ (DCIS). The mean $\pm$ standard deviation of the DSC, PR-AUC, 95HD, and 1mmCD across all test patients is shown per breast cancer (BCa) type, per model input, 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 BCa type and model input. Following Bonferroni-Holm correction ($\alpha = 0.05$, $m = 4$), no significant performance differences were found.

BCa type:		ILC (N=6)		
Metric	Input	Intens thresh	ResU-Net	p ResU-Net vs. Intens thresh
DSC	micro-CT	0.27 $\pm$ 0.24	0.27 $\pm$ 0.24	0.3125
PR-AUC	micro-CT	0.39 $\pm$ 0.26	0.31 $\pm$ 0.31	1.0000
95HD	micro-CT	6.47 $\pm$ 3.46	7.32 $\pm$ 2.81	1.0000
1mmCD	micro-CT	0.27 $\pm$ 0.07	0.29 $\pm$ 0.18	0.5625
DSC	micro-PET	0.33 $\pm$ 0.17	0.48 $\pm$ 0.24	0.1562
PR-AUC	micro-PET	0.47 $\pm$ 0.28	0.58 $\pm$ 0.20	0.0938
95HD	micro-PET	6.52 $\pm$ 4.21	5.30 $\pm$ 4.50	0.0938
1mmCD	micro-PET	0.31 $\pm$ 0.06	0.39 $\pm$ 0.11	0.5625
DSC	micro-PET-CT	0.41 $\pm$ 0.18	0.49 $\pm$ 0.26	0.2188
PR-AUC	micro-PET-CT	0.52 $\pm$ 0.29	0.60 $\pm$ 0.25	0.0625
95HD	micro-PET-CT	6.27 $\pm$ 4.28	5.88 $\pm$ 4.34	0.5625
1mmCD	micro-PET-CT	0.37 $\pm$ 0.15	0.46 $\pm$ 0.25	1.0000
BCa type:		DCIS (N=3)		
Metric	Input	Intens thresh	ResU-Net	p ResU-Net vs. Intens thresh
DSC	micro-CT	0.02 $\pm$ 0.00	0.02 $\pm$ 0.00	0.5000
PR-AUC	micro-CT	0.01 $\pm$ 0.00	0.01 $\pm$ 0.00	0.7500
95HD	micro-CT	15.97 $\pm$ 4.02	15.53 $\pm$ 4.49	0.5000
1mmCD	micro-CT	0.12 $\pm$ 0.05	0.18 $\pm$ 0.07	0.2500
DSC	micro-PET	0.02 $\pm$ 0.02	0.03 $\pm$ 0.02	0.5000
PR-AUC	micro-PET	0.02 $\pm$ 0.01	0.17 $\pm$ 0.15	0.2500
95HD	micro-PET	12.13 $\pm$ 6.08	12.50 $\pm$ 6.23	0.2500
1mmCD	micro-PET	0.31 $\pm$ 0.17	0.27 $\pm$ 0.13	0.5000
DSC	micro-PET-CT	0.02 $\pm$ 0.02	0.03 $\pm$ 0.02	0.5000
PR-AUC	micro-PET-CT	0.02 $\pm$ 0.01	0.03 $\pm$ 0.03	1.0000
95HD	micro-PET-CT	12.96 $\pm$ 6.14	12.68 $\pm$ 6.47	1.0000
1mmCD	micro-PET-CT	0.23 $\pm$ 0.09	0.33 $\pm$ 0.19	0.5000

DSC = Dice similarity coefficient, PR-AUC = area under the curve of the precision-recall curve, 95HD = 95th percentile of the Hausdorff distance (in mm), 1mmCD = contour Dice with a 1 mm tolerance

Table D.2. P-values (p) resulting from the Wilcoxon signed-rank test, comparing the performance of the ResU-Nets and intensity thresholding (Intens thresh) models across different model inputs for segmentation of invasive lobular carcinoma (ILC) and ductal carcinoma in situ (DCIS). P-values are reported for each metric and each comparison of model inputs, stratified per breast cancer (BCa) type and segmentation strategy. Following Bonferroni-Holm correction ($\alpha = 0.05$, $m = 12$), no significant performance differences were found.

BCa type:		ILC (N=6)		
Metric	Model	p micro-PET vs. micro-CT	p micro-PET-CT vs. micro-CT	p micro-PET-CT vs. micro-PET
DSC	Intens thresh	0.4375	0.1562	0.0312
PR-AUC	Intens thresh	0.0938	0.0625	0.0312
95HD	Intens thresh	0.8438	1.0000	0.0625
1mmCD	Intens thresh	0.3125	0.1562	0.2188
DSC	ResU-Net	0.1562	0.1562	1.0000
PR-AUC	ResU-Net	0.0625	0.0312	0.5625
95HD	ResU-Net	0.4375	0.6250	0.8438
1mmCD	ResU-Net	0.5625	0.5625	0.4375
BCa type:		DCIS (N=3)		
Metric	Model	p micro-PET vs. micro-CT	p micro-PET-CT vs. micro-CT	p micro-PET-CT vs. micro-PET
DSC	Intens thresh	0.7500	0.7500	1.0000
PR-AUC	Intens thresh	0.2500	0.2500	0.5000
95HD	Intens thresh	0.2500	0.5000	0.2500
1mmCD	Intens thresh	0.2500	0.2500	0.7500
DSC	ResU-Net	0.7500	0.7500	0.7500
PR-AUC	ResU-Net	0.2500	0.5000	0.2500
95HD	ResU-Net	0.2500	0.5000	0.5000
1mmCD	ResU-Net	0.2500	0.2500	0.2500

DSC = Dice similarity coefficient, PR-AUC = area under the curve of the precision-recall curve, 95HD = 95th percentile of the Hausdorff distance (in mm), 1mmCD = contour Dice with a 1 mm tolerance

E | Margin assessment performance for micro-PET-CT images of ILC, DCIS, and neoadjuvant-treated lumpectomy specimens

This appendix contains the detailed evaluation of the margin assessment performance of the Residual U-Nets (ResU-Nets) and intensity thresholding models when applied to micro-PET-CT images of lumpectomy specimens with invasive lobular carcinoma (ILC) and ductal carcinoma in situ (DCIS), as well as of neoadjuvant-treated (NAT) specimens. These NAT specimens include invasive carcinomas of no special type (NST) pretreated with neoadjuvant chemotherapy (NAC), and cases of NST pretreated with aromatase inhibitors (NST-AI) and ILC pretreated with aromatase inhibitors (ILC-AI). Margin assessment performance is evaluated separately for the datasets collected in the BIMAP and BrIMA studies.

E.1. BIMAP

Visualisations of predicted tumour contours for two ILC, two DCIS, and two NAC specimens are shown in Figures E.1, E.2, and E.3, respectively. Figure E.4 shows the contingency tables for the ILC, DCIS, and NAC cases, comparing the predicted and true histopathological margin status. Separate tables were constructed for the predictions obtained using the ResU-Nets and intensity thresholding models, as well as for predictions based on the postoperative majority vote of seven physicians. From these contingency tables, sensitivity, specificity, precision, and F1 score for positive margin detection were calculated for each margin assessment strategy. Table E.1 shows all metric values for the ILC, DCIS, and NAC cases. P-values resulting from McNemar's test to evaluate differences in margin assessment performance between the different strategies are reported in Figure E.5 for the ILC, DCIS, and NAC cases.

For the ILC cases, the best performance was reached by the postoperative majority vote of seven physicians, followed by the ResU-Nets, and intensity thresholding models. For the DCIS cases, the margin status was correctly predicted for the two available specimens using all three methods. For the NAC cases, the ResU-Nets and intensity thresholding models both classified one of two cases correctly, while the postoperative majority vote of seven physicians lead to two incorrect predictions. When studying the ILC, DCIS, and NAC cases separately, none of the differences between the three margin assessment strategies were found to be statistically significant.

Table E.1. Metric values to evaluate the agreement between the true histopathological and predicted margin status for the BIMAP dataset, shown separately for invasive lobular carcinoma (ILC), for ductal carcinoma in situ (DCIS), and for invasive carcinoma of no special type pretreated with neoadjuvant chemotherapy (NAC). 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 postoperative majority vote of seven physicians. The performance of these seven physicians was previously described by De Crem et al. [32].

BCa type:		ILC (N=6)	
Metrics	Intens thresh	ResU-Net	7 Physicians Postoperative
Sensitivity (Recall)	0.50	0.50	0.50
Specificity	0.25	0.50	0.75
Precision	0.25	0.33	0.50
F1 score	0.33	0.40	0.50
BCa type:		DCIS (N=2)	
Metrics	Intens thresh	ResU-Net	7 Physicians Postoperative
Sensitivity (Recall)	1.00	1.00	1.00
Specificity	1.00	1.00	1.00
Precision	1.00	1.00	1.00
F1 score	1.00	1.00	1.00
BCa type:		NAC (N=2)	
Metrics	Intens thresh	ResU-Net	7 Physicians Postoperative
Sensitivity (Recall)	1.00	1.00	0.00
Specificity	0.00	0.00	0.00
Precision	0.50	0.50	0.00
F1 score	0.67	0.67	NA

NA = not applicable

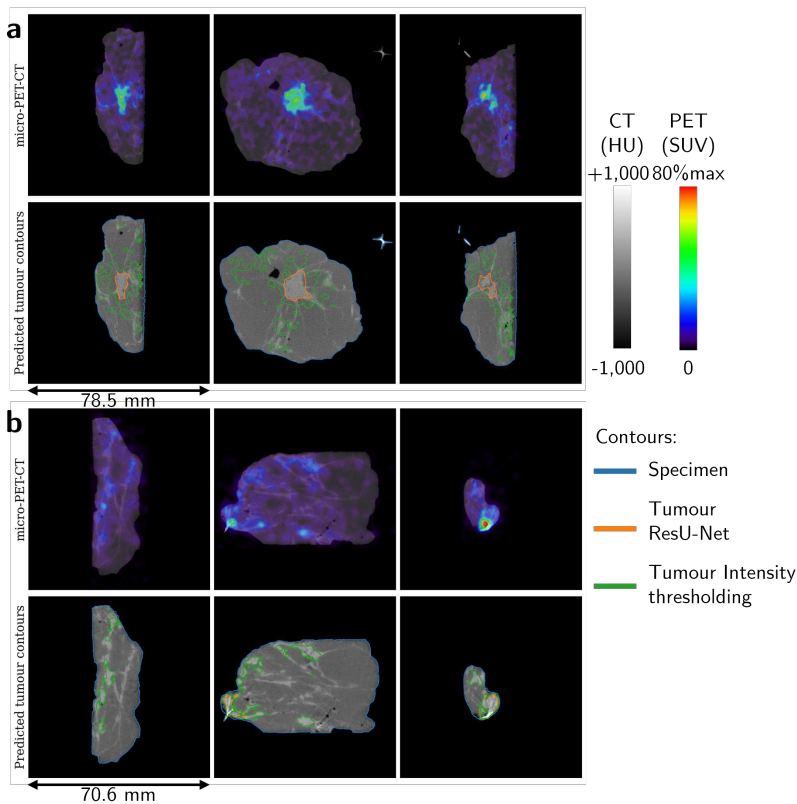


Figure E.1. Predicted tumour contours for two (a and b) invasive lobular carcinomas from the BIMAP dataset. For every specimen, three orthogonal slices are visualised. The top row shows the micro-PET-CT input (expressed in standardised uptake values (SUVs) and Hounsfield units (HUs)), and the bottom row shows the contours overlaid on the micro-CT (blue = specimen, orange = tumour predicted by the Residual U-Net (ResU-Net), green = tumour predicted through intensity thresholding). (a) The ResU-Net prediction is true negative, while the intensity thresholding prediction is false positive. (b) Both ResU-Net and intensity thresholding predictions are true positive.

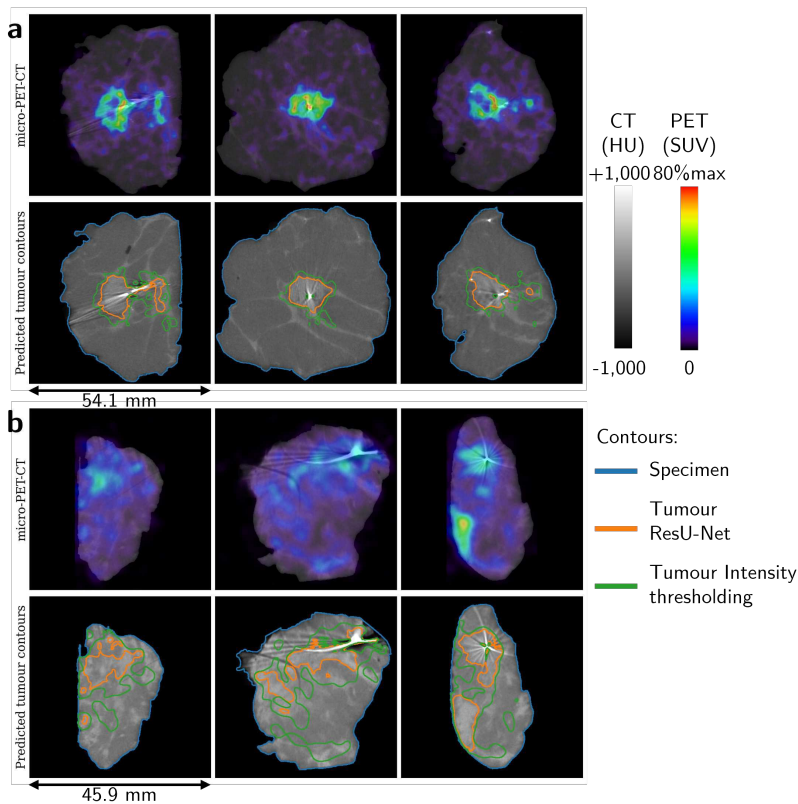


Figure E.2. Predicted tumour contours for two (a and b) ductal carcinoma in situ specimens from the BIMAP dataset. For every specimen, three orthogonal slices are visualised. The top row shows the micro-PET-CT input (expressed in standardised uptake values (SUVs) and Hounsfield units (HUs)), and the bottom row shows the contours overlaid on the micro-CT (blue = specimen, orange = tumour predicted by the Residual U-Net (ResU-Net), green = tumour predicted through intensity thresholding). (a) Both ResU-Net and intensity thresholding predictions are true negative. (b) Both ResU-Net and intensity thresholding predictions are true positive.

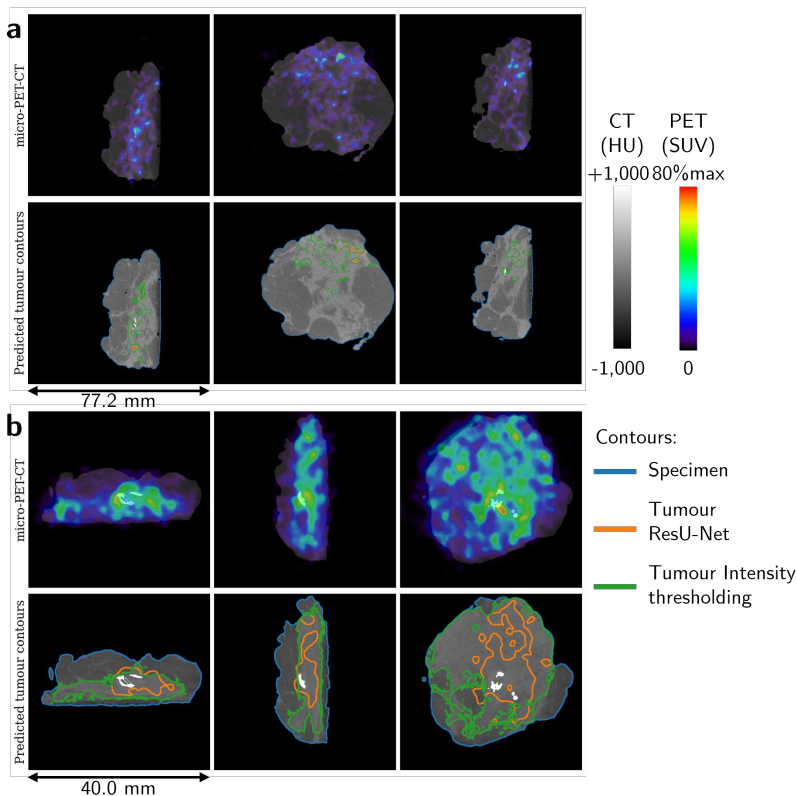


Figure E.3. Predicted tumour contours for two (a and b) invasive carcinomas of no special type from the BIMAP dataset, pretreated with neoadjuvant chemotherapy. For every specimen, three orthogonal slices are visualised. The top row shows the micro-PET-CT input (expressed in standardised uptake values (SUVs) and Hounsfield units (HUs)), and the bottom row shows the contours overlaid on the micro-CT (blue = specimen, orange = tumour predicted by the Residual U-Net (ResU-Net), green = tumour predicted through intensity thresholding). (a) Both ResU-Net and intensity thresholding predictions are false positive. (b) Both ResU-Net and intensity thresholding predictions are true positive.

		Prediction			Intens thresh		ResU-Net		7 Physicians Postoperative	
		+	-							
Ground Truth	+	TP	FN	ILC (N=6)	1	1	1	1	1	1
	-	FP	TN		3	1	2	2	1	3
				DCIS (N=2)	1	0	1	0	1	0
					0	1	0	1	0	1
				NAC (N=2)	1	0	1	0	0	1
					1	0	1	0	1	0

Figure E.4. Contingency tables for margin status prediction in the BIMAP dataset, shown separately for invasive lobular carcinoma (ILC), for ductal carcinoma in situ (DCIS), and for invasive carcinoma of no special type pretreated with neoadjuvant chemotherapy (NAC). Margin status was predicted using intensity thresholding (Intens thresh) and the Residual U-Net (ResU-Net) applied to the micro-PET-CT input, as well as through the postoperative majority vote of seven physicians. The predicted margin status was evaluated against the true histopathological margin status. The performance of the seven physicians was previously described by De Crem et al. [32].

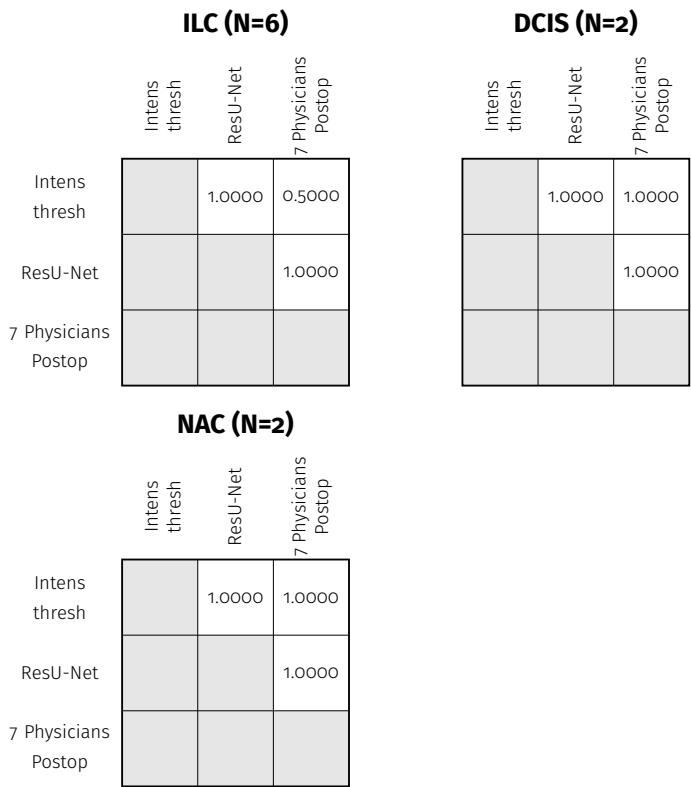


Figure E.5. P-values (p) resulting from McNemar’s test to compare different approaches to predict the margin status for the BIMAP dataset, shown separately for invasive lobular carcinoma (ILC), for ductal carcinoma in situ (DCIS), and for invasive carcinoma of no special type pretreated with neoadjuvant chemotherapy (NAC). The following strategies were compared: postoperative majority voting of seven physicians, prediction with Residual U-Nets (ResU-Nets), and prediction with intensity thresholding (Intens thresh) models. Following Bonferroni-Holm correction ($\alpha = 0.05$, $m = 3$), no significant performance differences were found.

E.2. BrIMA

Visualisations of predicted tumour contours for two ILC, two DCIS, and three NAT specimens (one NST after NAC, one NST-AI, and one ILC-AI) are shown in Figures E.6, E.7, and E.8, respectively. Figure E.9 shows the contingency tables for the ILC, DCIS, and combined NAT cases. Separate tables were constructed for the predictions obtained using the ResU-Nets and intensity thresholding models, as well as for the physicians' intraoperative predictions and the postoperative majority vote of four physicians. Table E.2 shows the sensitivity, specificity, precision, and F1 score for positive margin detection for each margin assessment strategy for the ILC, DCIS, and combined NAT cases. P-values resulting from McNemar's test to evaluate differences in margin assessment performance between the different strategies are reported in Figure E.10 for the ILC, DCIS, and combined NAT cases.

For the ILC cases, the sensitivity of the ResU-Nets was lower, while the specificity, precision, and F1 score were higher compared to the intraoperative physicians' predictions and the postoperative majority vote of four physicians. The ResU-Net predictions were found to be statistically equivalent to the intraoperative and postoperative physicians' predictions. The performance of the intensity thresholding models was found to be significantly different from the performance of the ResU-Nets and the intraoperative predictions of the physicians, with a lower specificity, precision, and F1 score. For the DCIS cases, none of the differences between the four margin assessment strategies were found to be statistically significant. The specificity was 0% for all strategies, except for the intraoperative interpretations of the physicians where two out of three specimens with a negative margin were correctly identified. For the NAT cases, the ResU-Net predictions were found to be statistically equivalent to both intraoperative and postoperative predictions of the physicians. The performance of the intensity thresholding models was found to be significantly different from the performance of the ResU-Nets and the intraoperative predictions of the physicians, with a lower specificity, precision, and F1 score. For the NAT cases, a significant difference was also found between the intraoperative physicians' predictions and the postoperative majority vote of four physicians.

Table E.2. Metric values to evaluate the agreement between the true histopathological and predicted margin status for the BrIMA dataset, shown separately for invasive lobular carcinoma (ILC), for ductal carcinoma in situ (DCIS), and for all neoadjuvant-treated (NAT) cases combined. The combined group of NAT cases consists of invasive carcinomas of no special type (NST) pretreated with neoadjuvant chemotherapy (NAC), cases of NST pretreated with aromatase inhibitors (NST-AI), and cases of ILC pretreated with aromatase inhibitors (ILC-AI). 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 intraoperative interpretations of the physicians, as well as to the postoperative majority vote of four physicians. The performance of the physicians is also described by Göker et al. [33].

BCa type:		ILC (N=16)		
Metrics	Intens thresh	ResU-Net	Physicians Intraoperative	4 Physicians Postoperative
Sensitivity (Recall)	1.00	0.67	0.83	1.00
Specificity	0.00	0.80	0.70	0.30
Precision	0.38	0.67	0.63	0.46
F1 score	0.55	0.67	0.71	0.63
BCa type:		DCIS (N=13)		
Metrics	Intens thresh	ResU-Net	Physicians Intraoperative	4 Physicians Postoperative
Sensitivity (Recall)	1.00	0.90	0.80	0.80
Specificity	0.00	0.00	0.67	0.00
Precision	0.77	0.75	0.89	0.73
F1 score	0.87	0.82	0.84	0.76
BCa type:		NAT (N=39)		
Metrics	Intens thresh	ResU-Net	Physicians Intraoperative	4 Physicians Postoperative
Sensitivity (Recall)	1.00	0.80	0.80	1.00
Specificity	0.24	0.59	0.68	0.41
Precision	0.16	0.22	0.27	0.20
F1 score	0.28	0.35	0.40	0.33

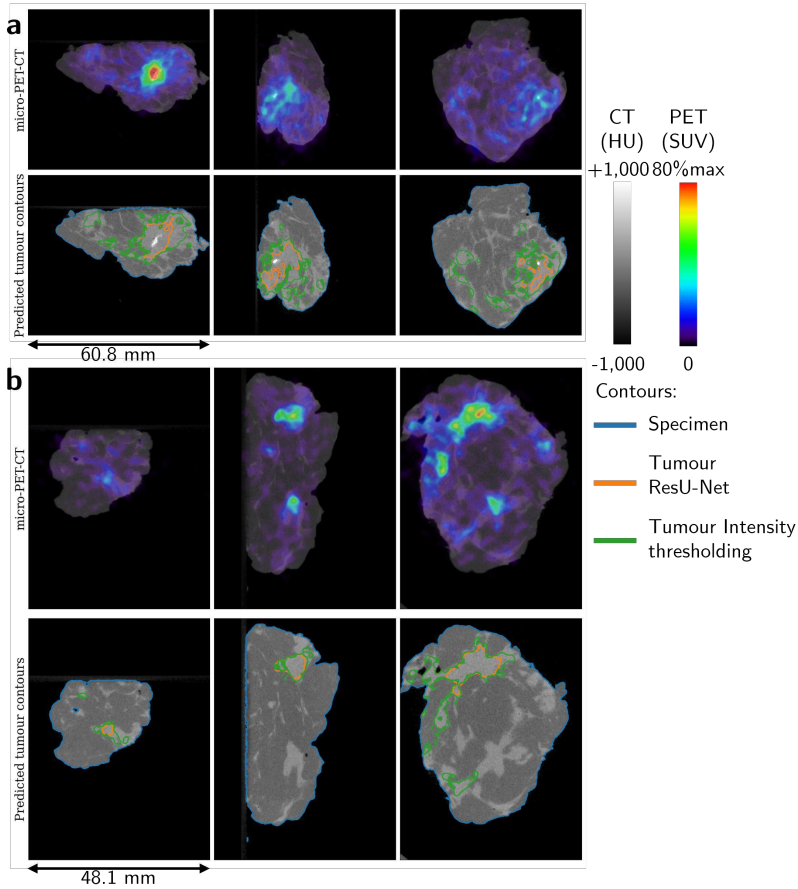


Figure E.6. Predicted tumour contours for two (a and b) invasive lobular carcinomas from the BrIMA dataset. For every specimen, three orthogonal slices are visualised. The top row shows the micro-PET-CT input (expressed in standardised uptake values (SUVs) and Hounsfield units (HUs)), and the bottom row shows the contours overlaid on the micro-CT (blue = specimen, orange = tumour predicted by the Residual U-Net (ResU-Net), green = tumour predicted through intensity thresholding). (a) The ResU-Net prediction is true negative, while the intensity thresholding prediction is false positive. (b) The ResU-Net prediction is false negative, while the intensity thresholding prediction is true positive.

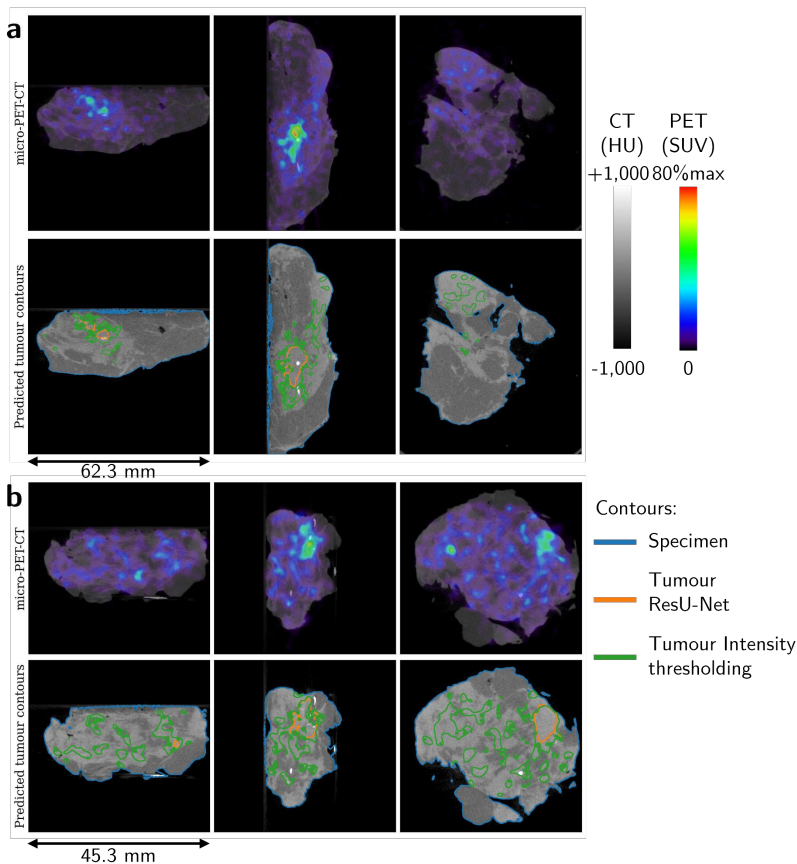


Figure E.7. Predicted tumour contours for two (a and b) ductal carcinoma in situ specimens from the BrIMA dataset. For every specimen, three orthogonal slices are visualised. The top row shows the micro-PET-CT input (expressed in standardised uptake values (SUVs) and Hounsfield units (HUs)), and the bottom row shows the contours overlaid on the micro-CT (blue = specimen, orange = tumour predicted by the ResU-Net, green = tumour predicted through intensity thresholding). (a) The ResU-Net prediction is false negative, while the intensity thresholding prediction is true positive. (b) Both ResU-Net and intensity thresholding predictions are false positive.

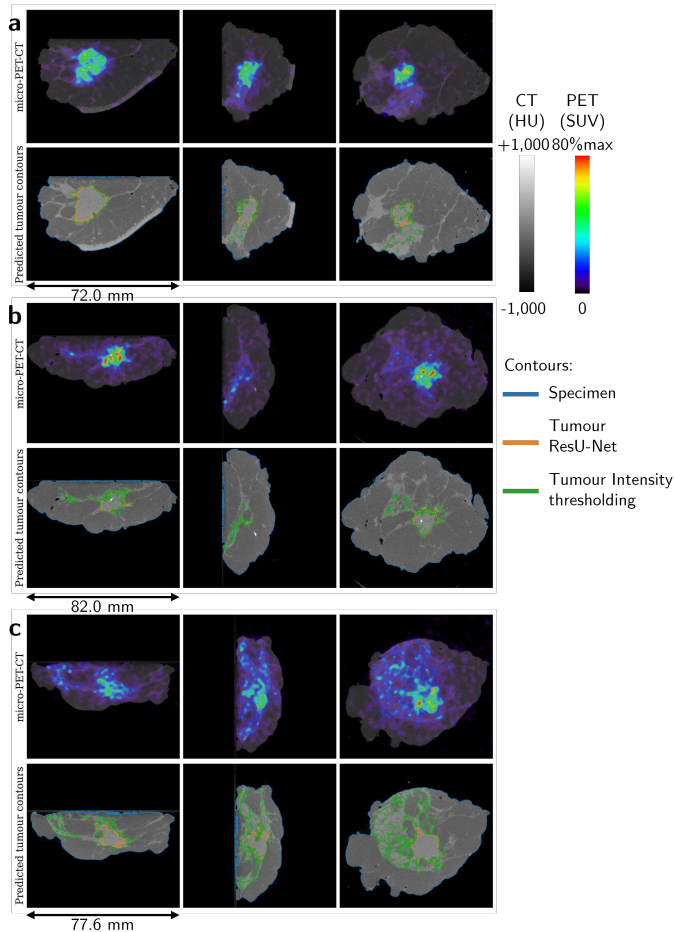


Figure E.8. Predicted tumour contours for three (a, b, and c) neoadjuvant-treated (NAT) lumpectomy specimens from the BrIMA dataset. For every specimen, three orthogonal slices are visualised. The top row shows the micro-PET-CT input (expressed in standardised uptake values (SUVs) and Hounsfield units (HUs)), and the bottom row shows the contours overlaid on the micro-CT (blue = specimen, orange = tumour predicted by the Residual U-Net (ResU-Net), green = tumour predicted through intensity thresholding). (a) Invasive carcinoma of no special type (NST) pretreated with neoadjuvant chemotherapy. The ResU-Net prediction is true negative, while the intensity thresholding prediction is false positive. (b) NST case pretreated with aromatase inhibitors (NST-AI). The ResU-Net prediction is true negative, while the intensity thresholding prediction is false positive. (c) Invasive lobular carcinoma (ILC) pretreated with aromatase inhibitors (ILC-AI). The ResU-Net prediction is false negative, while the intensity thresholding prediction is true positive.

		Prediction											
		+	-										
Ground Truth	+	TP	FN	ILC (N=16)	Intens thresh			ResU-Net		Physicians Intraoperative		4 Physicians Postoperative	
	-	FP	TN		6	0	4	2	5	1	6	0	
					10	0	2	8	3	7	7	3	
				DCIS (N=13)	10	0	9	1	8	2	8	2	
					3	0	3	0	1	2	3	0	
				NAT (N=39)	5	0	4	1	4	1	5	0	
					26	8	14	20	11	23	20	14	

Figure E.9. Contingency tables for margin status prediction in the BrIMA dataset, shown separately for invasive lobular carcinoma (ILC), for ductal carcinoma in situ (DCIS), and for all neoadjuvant-treated (NAT) cases combined. The combined group of NAT cases consists of invasive carcinomas of no special type (NST) pretreated with neoadjuvant chemotherapy (NAC), cases of NST pretreated with aromatase inhibitors (NST-AI), and cases of ILC pretreated with aromatase inhibitors (ILC-AI). Margin status was predicted using intensity thresholding (Intens thresh) and the Residual U-Net (ResU-Net) applied to the micro-PET-CT input, as well as through intraoperative physicians' assessments and the postoperative majority vote of four physicians. The predicted margin status was evaluated against the true histopathological margin status. The performance of the physicians is also described by Göker et al. [33].

ILC (N=16)					DCIS (N=13)				
	Intens thresh	ResU-Net	Physicians Intraop	4 Physicians Postop		Intens thresh	ResU-Net	Physicians Intraop	4 Physicians Postop
Intens thresh		0.0020	0.0078	0.2500			1.0000	0.1250	0.5000
ResU-Net			0.5000	0.0156				0.2500	1.0000
Physicians Intraop				0.1250					0.6250
4 Physicians Postop									

NAT (N=39)				
	Intens thresh	ResU-Net	Physicians Intraop	4 Physicians Postop
Intens thresh		0.0002	0.0000	0.1094
ResU-Net			0.4531	0.0391
Physicians Intraop				0.0020
4 Physicians Postop				

Figure E.10. P-values (p) resulting from McNemar’s test to compare different approaches to predict the margin status for the BrIMA dataset, shown separately for invasive lobular carcinoma (ILC), for ductal carcinoma in situ (DCIS), and for all neoadjuvant-treated (NAT) cases combined. The combined group of NAT cases consists of invasive carcinomas of no special type (NST) pretreated with neoadjuvant chemotherapy (NAC), cases of NST pretreated with aromatase inhibitors (NST-AI), and cases of ILC pretreated with aromatase inhibitors (ILC-AI). The following strategies were compared: intraoperative interpretations of physicians, postoperative majority voting of four physicians, prediction with Residual U-Nets (ResU-Nets), and prediction with intensity thresholding (Intens thresh) models. Significant performance differences following Bonferroni-Holm correction ($\alpha = 0.05$, $m = 6$) are indicated in bold.

F | Leave-one-out strategy to choose an operating point for margin status prediction from micro-PET-CT images of lumpectomy specimens

In this appendix, we explore an alternative strategy for selecting an operating point on the receiver operating characteristic (ROC) and precision-recall (PR) curves to evaluate the margin assessment performance of the Residual U-Nets (ResU-Nets) and intensity thresholding models. In Section 6.3.1 of Chapter 6, we evaluated the performance of both models at the operating point corresponding to optimal segmentation performance as measured by the Dice similarity coefficient (DSC), hereafter referred to as the “segmentation operating point”. In this appendix, we instead assess the performance at an operating point chosen through a leave-one-out (LOO) strategy. For each case in the datasets collected in the BIMAP and BrIMA studies, the ROC- and PR-curves were constructed using all remaining cases in the dataset. For each case, the operating point was then selected as the highest threshold for which the sensitivity was at least equal to that achieved by the physicians. This procedure was performed for both the ResU-Nets and intensity thresholding models to choose an appropriate tumour probability threshold and a relative PET intensity threshold, respectively. Finally, for each case, the found LOO thresholds were applied to segment the tumour and predict the margin status.

The LOO strategy was applied separately for the BIMAP and BrIMA datasets. For the BIMAP dataset, we chose to target the sensitivity of the postoperative majority vote of seven physicians, whereas for the BrIMA dataset, we targeted the sensitivity of the physicians’ intraoperative predictions. For invasive carcinomas of no special type (NST), the operating points (i.e., the thresholds) were chosen based on all remaining NST cases in the same dataset. For invasive lobular carcinomas (ILCs) and cases of ductal carcinoma in situ (DCIS), the operating points were selected based on all remaining combined ILC and DCIS cases in the same dataset. For neoadjuvant-treated (NAT) cases, the operating points were chosen based on all remaining combined ILC, DCIS, and NAT cases within the same dataset. Results for the BIMAP and BrIMA datasets are shown in the sections below.

F.1. BIMAP

Figure F.1 displays the contingency tables, comparing the predicted and true histopathological margin status, for the NST cases and for all combined breast cancer (BCa) subtypes in the BIMAP dataset. Separate tables were constructed for the predictions obtained using the ResU-Nets and intensity thresholding models, as well as for predictions based on the postoperative majority vote of seven physicians. From these contingency tables, sensitivity, specificity, precision, and F1 score for positive margin detection were calculated for each margin assessment strategy. Table F.1 shows all metric values for the NST cases and for all combined BCa subtypes of the BIMAP dataset.

P-values resulting from McNemar’s test to evaluate differences in margin assessment performance between the ResU-Nets, intensity thresholding models, and the postoperative majority vote of seven physicians are reported in Figure F.2 for the NST cases and for all combined BCa subtypes in the BIMAP dataset. For the NST cases, the margin status predictions obtained with the ResU-Nets and intensity thresholding models were identical to those obtained through postoperative majority voting of seven physicians. For the combined group of all BCa subtypes in the BIMAP dataset, both the ResU-Nets and intensity thresholding models achieved a higher sensitivity than the postoperative majority vote of seven physicians, with a similar specificity and precision. As reported in Section 6.3.1.1 of Chapter 6 for the segmentation operating point, the margin assessment performance of the ResU-Nets was again found to be statistically equivalent to the majority vote of seven physicians, both for the NST cases and for the combined group of all BCa subtypes in the BIMAP dataset. In contrast to the findings reported in Section 6.3.1.1 of Chapter 6 for the segmentation operating point, no statistically significant differences in performance were here observed between the intensity thresholding models and either the ResU-Nets or the majority vote of seven physicians, both for the NST cases and the combined group of all BCa subtypes in the BIMAP dataset.

Table F.1. Metric values to evaluate the agreement between the true histopathological and predicted margin status for the BIMAP dataset, using the leave-one-out strategy to select the operating point, shown separately for invasive carcinomas of no special type (NST), and for all breast cancer (BCa) subtypes combined. 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 postoperative majority vote of seven physicians. The performance of these seven physicians was previously described by De Crem et al. [32].

BCa type:		NST (N=31)		
Metrics		Intens thresh	ResU-Net	7 Physicians Postoperative
Sensitivity (Recall)		0.89	0.89	0.89
Specificity		0.91	0.91	0.91
Precision		0.80	0.80	0.80
F1 score		0.84	0.84	0.84
BCa type:		All types (N=41)		
Metrics		Intens thresh	ResU-Net	7 Physicians Postoperative
Sensitivity (Recall)		0.85	0.85	0.77
Specificity		0.89	0.86	0.86
Precision		0.79	0.73	0.71
F1 score		0.81	0.79	0.74

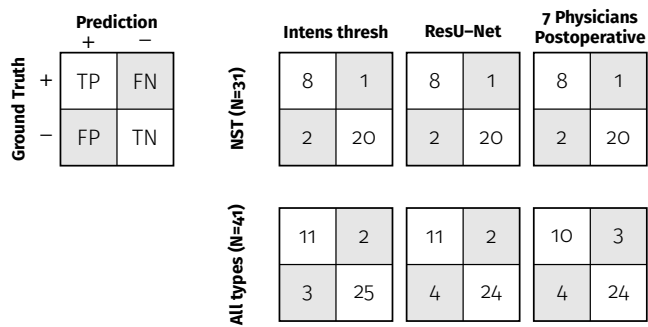


Figure F.1. Contingency tables for margin status prediction in the BIMAP dataset, using the leave-one-out strategy to select the operating point, separately for invasive carcinomas of no special type (NST), and for all breast cancer (BCa) subtypes combined. Margin status was predicted using intensity thresholding (Intens thresh) and the Residual U-Net (ResU-Net) applied to the micro-PET-CT input, as well as through the postoperative majority vote of seven physicians. The predicted margin status was evaluated against the true histopathological margin status. The performance of the seven physicians was previously described by De Crem et al. [32].

	NST (N=31)			All types (N=41)		
	Intens thresh	ResU-Net	7 Physicians Postop	Intens thresh	ResU-Net	7 Physicians Postop
Intens thresh		1.0000	1.0000		1.0000	1.0000
ResU-Net			1.0000			1.0000
7 Physicians Postop						

Figure F.2. P-values (p) resulting from McNemar’s test to compare different approaches to predict the margin status for the BIMAP dataset, using the leave-one-out strategy to select the operating point, shown separately for invasive carcinomas of no special type (NST), and for all breast cancer (BCa) subtypes combined. The following strategies were compared: postoperative majority voting of seven physicians, prediction with Residual U-Nets (ResU-Nets), and prediction with intensity thresholding (Intens thresh) models. Following Bonferroni-Holm correction ($\alpha = 0.05$, $m = 3$), no significant performance differences were found.

F.2. BrIMA

Figure F.3 displays the contingency results for the NST cases and all combined BCa subtypes in the BrIMA dataset. Separate tables were constructed for the predictions obtained using the ResU-Nets and intensity thresholding models, as well as for the physicians' intraoperative predictions and the postoperative majority vote of four physicians. From these contingency tables, sensitivity, specificity, precision, and F1 score for positive margin detection were calculated for each margin assessment strategy. Table F.2 shows all metric values for the NST cases and for all combined BCa subtypes of the BrIMA dataset.

P-values resulting from McNemar's test to evaluate differences in margin assessment performance between the ResU-Nets, intensity thresholding models, the intraoperative physicians' predictions, and the majority vote of four physicians, are reported in Figure F.4 for the NST cases and for all combined BCa subtypes in the BrIMA dataset. As reported in Section 6.3.1.2 of Chapter 6 for the segmentation operating point, the margin assessment performance of the ResU-Nets was again found to be statistically equivalent to the intraoperative physicians' predictions and to the majority vote of four physicians, both for the NST cases and for the combined group of all BCa subtypes in the BrIMA dataset. In agreement with the findings reported in Section 6.3.1.2 of Chapter 6 for the segmentation operating point, statistically significant differences in performance were observed between the intensity thresholding models and the ResU-Nets, both for the NST cases and the combined group of all BCa subtypes. In addition, the performance of the intensity thresholding models was again found to be significantly different from the intraoperative physicians' predictions and the postoperative majority vote of four physicians, both for the NST cases and the combined group of all BCa subtypes. The intensity thresholding models obtained a lower specificity, precision, and F1 score compared to the physicians' predictions. The consistently lower performance of the intensity thresholding models compared with the ResU-Nets for the BrIMA dataset, indicates that identifying a robust PET intensity threshold for accurate margin status prediction across a large number of cases remains challenging.

Table F.2. Metric values to evaluate the agreement between the true histopathological and predicted margin status for the BrIMA dataset, using the leave-one-out strategy to select the operating point, shown separately for invasive carcinomas of no special type (NST), and for all breast cancer (BCa) subtypes combined. 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 intraoperative interpretations of the physicians, as well as to the postoperative majority vote of four physicians. The performance of the physicians is also described by Göker et al. [33].

BCa type:		NST (N=84)		
Metrics	Intens thresh	ResU-Net	Physicians Intraoperative	4 Physicians Postoperative
Sensitivity (Recall)	0.86	0.71	0.71	0.71
Specificity	0.46	0.57	0.66	0.67
Precision	0.24	0.25	0.29	0.30
F1 score	0.38	0.37	0.42	0.43
BCa type:		All types (N=152)		
Metrics	Intens thresh	ResU-Net	Physicians Intraoperative	4 Physicians Postoperative
Sensitivity (Recall)	0.89	0.83	0.77	0.83
Specificity	0.42	0.54	0.67	0.55
Precision	0.31	0.35	0.41	0.35
F1 score	0.46	0.49	0.53	0.50

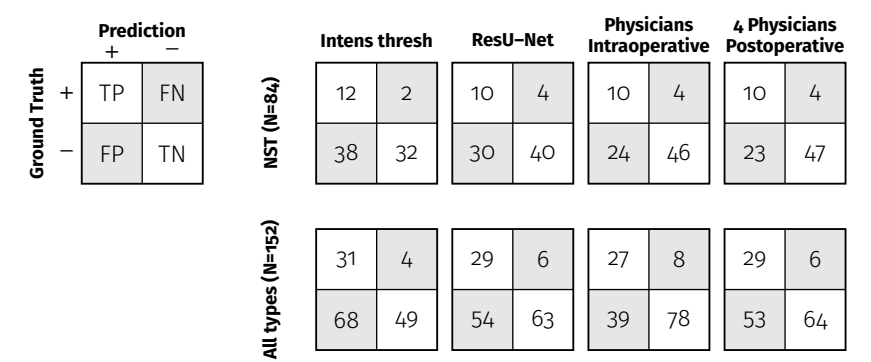


Figure F.3. Contingency tables for margin status prediction in the BrIMA dataset, using the leave-one-out strategy to select the operating point, shown separately for invasive carcinomas of no special type (NST), and for all breast cancer (BCa) subtypes combined. Margin status was predicted using intensity thresholding (Intens thresh) and the Residual U-Net (ResU-Net) applied to the micro-PET-CT input, as well as through intraoperative physicians' assessments and the postoperative majority vote of four physicians. The predicted margin status was evaluated against the true histopathological margin status. The performance of the physicians is also described by Göker et al. [33].

NST (N=84)					All types (N=152)				
	Intens thresh	ResU-Net	Physicians Intraop	4 Physicians Postop		Intens thresh	ResU-Net	Physicians Intraop	4 Physicians Postop
Intens thresh		0.0020	0.0015	0.0005			0.0001	0.0000	0.0046
ResU-Net			0.3075	0.1892				0.0095	1.0000
Physicians Intraop				1.0000					0.0259
4 Physicians Postop									

Figure F.4. P-values (p) resulting from McNemar’s test to compare different approaches to predict the margin status for the BrIMA dataset, using the leave-one-out strategy to select the operating point, shown separately for invasive carcinomas of no special type (NST), and for all breast cancer (BCa) subtypes combined. The following strategies were compared: intraoperative interpretations of physicians, postoperative majority voting of four physicians, prediction with Residual U-Nets (ResU-Nets), and prediction with intensity thresholding (Intens thresh) models. Significant performance differences following Bonferroni-Holm correction ($\alpha = 0.05$, $m = 6$) are indicated in bold.

Bibliography

- [1] Andrea M. Bodine, Brian Holahan, and Alyssa Mixon. Benign Breast Conditions. *Journal of Osteopathic Medicine*, 117(12):755–760, December 2017.
- [2] Louise Wilkinson and Toral Gathani. Understanding breast cancer as a global health concern. *The British Journal of Radiology*, 95(1130):20211033, February 2022.
- [3] Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249, May 2021.
- [4] Akram M. Shaaban. *Diagnostic imaging oncology*. Elsevier, 2020.
- [5] F. Cardoso, S. Kyriakides, S. Ohno, F. Penault-Llorca, P. Poortmans, I.T. Rubio, S. Zackrisson, and E. Senkus. Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of Oncology*, 30(8):1194–1220, August 2019.
- [6] Shuangjian Li, Xin Li, Dan Li, Qian Zhao, Liping Zhu, and Tao Wu. A meta-analysis of randomized controlled trials comparing breast-conserving surgery and mastectomy in terms of patient survival rate and quality of life in breast cancer. *International Journal for Quality in Health Care: Journal of the International Society for Quality in Health Care*, 36(2):mzaeo43, May 2024.
- [7] Caroline Koopmansch, Jean-Christophe Noël, Calliope Maris, Philippe Simon, Marième Sy, and Xavier Catteau. Intraoperative Evaluation of Resection Margins in Breast-Conserving Surgery for *In Situ* and Invasive Breast Carcinoma. *Breast Cancer: Basic and Clinical Research*, 15:117822342199345, January 2021.
- [8] Edward Robert St John, Rashed Al-Khudairi, Hutan Ashrafian, Thanos Athanasiou, Zoltan Takats, Dimitri John Hadjiminias, Ara Darzi, and Daniel Richard Leff. Diagnostic Accuracy of Intraoperative Techniques for Margin Assessment in Breast Cancer Surgery: A Meta-analysis. *Annals of Surgery*, 265(2):300–310, February 2017.
- [9] Bernard Fisher, Stewart Anderson, John Bryant, Richard G. Margoese, Melvin Deutsch, Edwin R. Fisher, Jong-Hyeon Jeong, and Norman Wolmark. Twenty-year follow-up of a randomized trial comparing total mastectomy, lumpectomy, and lumpectomy plus irradiation for the treatment of invasive breast cancer. *The New England Journal of Medicine*, 347(16):1233–1241, October 2002.

- [10] Matthew M. Poggi, David N. Danforth, Linda C. Sciuto, Sharon L. Smith, Seth M. Steinberg, David J. Liewehr, Cynthia Menard, Marc E. Lippman, Allen S. Lichter, and Rosemary M. Altamus. Eighteen-year results in the treatment of early breast carcinoma with mastectomy versus breast conservation therapy: The National Cancer Institute randomized trial. *Cancer*, 98(4):697–702, August 2003.
- [11] Emil D. Kurniawan, Matthew H. Wong, Imogen Windle, Allison Rose, Arlene Mou, Malcolm Buchanan, John P. Collins, Julie A. Miller, Russell L. Gruen, and G. Bruce Mann. Predictors of Surgical Margin Status in Breast-Conserving Surgery Within a Breast Screening Program. *Annals of Surgical Oncology*, 15(9):2542–2549, September 2008.
- [12] Umberto Veronesi, Natale Cascinelli, Luigi Mariani, Marco Greco, Roberto Saccozzi, Alberto Luini, Marisel Aguilar, and Ettore Marubini. Twenty-year follow-up of a randomized study comparing breast-conserving surgery with radical mastectomy for early breast cancer. *The New England Journal of Medicine*, 347(16):1227–1232, October 2002.
- [13] S.Eva Singletary. Surgical margins in patients with early-stage breast cancer treated with breast conservation therapy. *The American Journal of Surgery*, 184(5):383–393, November 2002.
- [14] M. Keskek, M. Kothari, B. Ardehali, N. Betambeau, N. Nasiri, and G.P.H. Gui. Factors predisposing to cavity margin positivity following conservation surgery for breast cancer. *European Journal of Surgical Oncology (EJSO)*, 30(10):1058–1064, December 2004.
- [15] Elvira L. Vos, Agnes Jager, Cornelis Verhoef, Adri C. Voogd, and Linetta B. Koppert. Overall survival in patients with a re-excision following breast conserving surgery compared to those without in a large population-based cohort. *European Journal of Cancer*, 51(3):282–291, February 2015.
- [16] Stewart J. Anderson, Irene Wapnir, James J. Dignam, Bernard Fisher, Eleftherios P. Mamounas, Jong-Hyeon Jeong, Charles E. Geyer, D. Lawrence Wickerham, Joseph P. Costantino, and Norman Wolmark. Prognosis after ipsilateral breast tumor recurrence and locoregional recurrences in patients treated by breast-conserving therapy in five National Surgical Adjuvant Breast and Bowel Project protocols of node-negative breast cancer. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 27(15):2466–2473, May 2009.
- [17] Liska Havel, Himani Naik, Luis Ramirez, Monica Morrow, and Jeffrey Landercasper. Impact of the SSO-ASTRO Margin Guideline on Rates of Re-excision After Lumpectomy for Breast Cancer: A Meta-analysis. *Annals of Surgical Oncology*, 26(5):1238–1244, May 2019.

- [18] Meena S. Moran, Stuart J. Schnitt, Armando E. Giuliano, Jay R. Harris, Seema A. Khan, Janet Horton, Suzanne Klimberg, Mariana Chavez-MacGregor, Gary Freedman, Nehmat Houssami, Peggy L. Johnson, and Monica Morrow. 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. *International Journal of Radiation Oncology*Biology*Physics*, 88(3):553–564, March 2014.
- [19] Jennifer F. Waljee, Emily S. Hu, Lisa A. Newman, and Amy K. Alderman. Predictors of re-excision among women undergoing breast-conserving surgery for cancer. *Annals of Surgical Oncology*, 15(5):1297–1303, May 2008.
- [20] Elizabeth S. McDonald, Amy S. Clark, Julia Tchou, Paul Zhang, and Gary M. Freedman. Clinical Diagnosis and Management of Breast Cancer. *Journal of Nuclear Medicine*, 57(Supplement 1):9S–16S, February 2016.
- [21] Susan G. Brouwer de Koning, Marie-Jeanne T. F. D. Vrancken Peeters, Katarzyna Jóźwiak, Patrick A. Bhairosing, and Theo J. M. Ruers. Tumor Resection Margin Definitions in Breast-Conserving Surgery: Systematic Review and Meta-analysis of the Current Literature. *Clinical Breast Cancer*, 18(4):e595–e600, August 2018.
- [22] James R. Bundred, Sarah Michael, Beth Stuart, Ramsey I. Cutress, Kerri Beckmann, Bernd Holleczeck, Jane E. Dahlstrom, Jacqui Gath, David Dodwell, and Nigel J. Bundred. Margin status and survival outcomes after breast cancer conservation surgery: prospectively registered systematic review and meta-analysis. *BMJ (Clinical research ed.)*, 378:e070346, September 2022.
- [23] Y. Grant, R. Al-Khudairi, E. St John, M. Barschkett, D. Cunningham, R. Al-Mufti, K. Hogben, P. Thiruchelvam, D. J. Hadjiminias, A. Darzi, A. W. Carter, and D. R. Leff. Patient-level costs in margin re-excision for breast-conserving surgery. *The British Journal of Surgery*, 106(4):384–394, March 2019.
- [24] Shoko E. Abe, Joshua S. Hill, Yimei Han, Kendall Walsh, James T. Symanowski, Lejla Hadzikadic-Gusic, Teresa Flippo-Morton, Terry Sarantou, Meghan Forster, and Richard L. White. Margin re-excision and local recurrence in invasive breast cancer: A cost analysis using a decision tree model. *Journal of Surgical Oncology*, 112(4):443–448, September 2015.
- [25] Jeffrey Landercasper, Eric Whitacre, Amy C. Degnim, and Mohammed Al-Hamadani. Reasons for Re-Excision After Lumpectomy for Breast Cancer: Insight from the American Society of Breast Surgeons MasterySM Database. *Annals of Surgical Oncology*, 21(10):3185–3191, October 2014.

- [26] Rares Georgescu, Flavian Tutuianu, Orsolya Bauer, Anca Toganel, Zalan Benedek, Eugeniu Darii, Sabin Turdean, and Cristina Tutuianu Radoi. Risk Factors for Positive Resection Margins in Breast-Conserving Surgery for Breast Cancer-Retrospective Analysis. *Cancers*, 16(17):2930, August 2024.
- [27] Julie A. Margenthaler and Aislinn Vaughan. Breast Conservation Surgery and the Definition of Adequate Margins: More Is Not Better... It's Just More. *JAMA Surgery*, 149(12):1305, December 2014.
- [28] Benjamin W. Maloney, David M. McClatchy, Brian W. Pogue, Keith D. Paulsen, Wendy A. Wells, and Richard J. Barth. Review of methods for intraoperative margin detection for breast conserving surgery. *Journal of Biomedical Optics*, 23(10):1, October 2018.
- [29] Sarah Goldfeder, Delphine Davis, and Jeanne Cullinan. Breast Specimen Radiography. *Academic Radiology*, 13(12):1453–1459, December 2006.
- [30] Menekse Göker, Radosław Marcinkowski, Mieke Van Bockstal, Vincent Keereman, Roel Van Holen, Jo Van Dorpe, Stefaan Vandenbergh, Boudewijn Brans, Herman Depypere, and Rudy Van den Broecke. 18F-FDG micro-PET/CT for intra-operative margin assessment during breast-conserving surgery. *Acta Chirurgica Belgica*, 120(5):366–374, 2020.
- [31] Bieke Lambert, Valérie Vergucht, Sam Dekeyser, Annick De Craene, Filip Ameye, Bieke Van Den Bossche, Dieter Berwouts, Jeroen Mertens, Henk Vanoverschelde, Mieke Coppens, Charlotte Gabriel, Marianne Rottiers, Carole Van Haverbeke, Peter Dekuyper, Pieter De Backer, Kenneth Carels, Tessa Van Oostveldt, and Karel Decaestecker. Feasibility study on the implementation of a mobile high-resolution PET/CT scanner for surgical specimens: exploring clinical applications and practical considerations. *European Journal of Nuclear Medicine and Molecular Imaging*, February 2025.
- [32] Anne-Sofie De Crem, Philippe Tummers, Herman Depypere, Geert Braems, Rawand Salihi, Glenn Vergauwen, Giovanni Cisternino, Koen Van de Vijver, Pieter De Visschere, Kathia De Man, Bliede Van den Broeck, Sigi Hendrickx, Liv Veldeman, Christel Monten, Jens M. Debacker, Hannelore Denys, and Menekse Göker. Breast cancer Intraoperative Margin Assessment using specimen PET-CT (BIMAP). *npj Breast Cancer*, 11(1):101, September 2025.
- [33] Menekse Göker, Markus Hahn, Ximena Elzo Kraemer, Hakima Harrach, Marian Vanhoeij, Glenn Vergauwen, Koen Van de Vijver, Kathia De Man, Fabian Schmidt, Selin Guergan, Valérie Vergucht, Rosa Di Micco, Giovanni Cisternino, and Oreste Davide Gentilini. Specimen PET-CT Imaging for Intraoperative Margin Assessment in Early-Stage

- Breast Cancer: The BrIMA Nonrandomized Clinical Trial. *JAMA Surgery*, June 2026.
- [34] Menekse Göker, Glenn Vergauwen, Charles Van de Velde, Koen Van de Vijver, Leen Van de Sande, and Chris Vanhove. High-resolution 18F-FDG PET-CT for peri-operative margin assessment in early- stage breast cancer: results of an ongoing clinical study. *Abstract and Poster presentation - SABCS*, 2021.
 - [35] Alexandros Moraitis, Theresa Kahl, Jens Kandziora, Walter Jentzen, David Kersting, Lukas Püllen, Henning Reis, Jens Köllermann, Claudia Kesch, Ulrich Krafft, Boris A. Hadaschik, Habib Zaidi, Ken Herrmann, Francesco Barbato, Wolfgang P. Fendler, Christopher Darr, and Pedro Fragosso Costa. Evaluation of Surgical Margins with Intraoperative PSMA PET/CT and Their Prognostic Value in Radical Prostatectomy. *Journal of Nuclear Medicine*, page jnumed.124.268719, February 2025.
 - [36] Peter J Bazira. Anatomy and physiology of the breast. *Surgery (Oxford)*, 42(12):859–864, December 2024.
 - [37] Herman Depypere, Philippe De Sutter, Yves Jacquemyn, and Ignace Vergrote. *Handboek Gynaecologie*. D/2017/0543/278. Acco, 2021 edition, 2017.
 - [38] Niketa Chotai and Supriya Kulkarni. General Considerations in Breast Imaging. In Niketa Chotai and Supriya Kulkarni, editors, *Breast Imaging Essentials: Case Based Review*, pages 3–8. Springer, Singapore, 2020.
 - [39] PDQ Screening and Prevention Editorial Board. Breast Cancer Screening (PDQ®): Patient Version. In *PDQ Cancer Information Summaries*. National Cancer Institute (US), Bethesda (MD), 2002.
 - [40] Sarah E Pinder and Ian O Ellis. The diagnosis and management of pre-invasive breast disease: Ductal carcinoma in situ (DCIS) and atypical ductal hyperplasia (ADH) – current definitions and classification. *Breast Cancer Research : BCR*, 5(5):254–257, 2003.
 - [41] Shadi Azam, Mikael Eriksson, Arvid Sjölander, Marike Gabrielson, Roxanna Hellgren, Kamila Czene, and Per Hall. Mammographic microcalcifications and risk of breast cancer. *British Journal of Cancer*, 125(5):759–765, August 2021.
 - [42] Giovanni Corso, Nicola Fusco, Elena Guerini-Rocco, Maria Cristina Leonardi, Carmen Criscitiello, Paola Zagami, Eleonora Nicolò, Giovanni Mazzarol, Carlo La Vecchia, Filippo Pesapane, Cristina Zanzottera, Paolo Tarantino, Salvatore Petitto, Beatrice Bianchi, Giulia Massari, Anthony Boato, Andrea Sibilio, Andrea Polizzi, Giuseppe Curigliano, Alessandra Margherita De Scalzi, Federica Lauria, Bernardo Bonanni, Monica Marabelli, Anna Rotili, Luca

- Nicosia, Adriana Albin, Mariarosaria Calvello, Rita A. Mukhtar, Mark E. Robson, Virgilio Sacchini, Gad Rennert, Viviana Galimberti, Paolo Veronesi, and Francesca Magnoni. Invasive lobular breast cancer: Focus on prevention, genetics, diagnosis, and treatment. *Seminars in Oncology*, 51(3-4):106–122, 2024.
- [43] Jill Seladi-Schulman. Invasive Breast Cancer: Types, Symptoms, Treatment, and Outlook, April 2021. Healthline. Available from <https://www.healthline.com/health/breast-cancer/invasive-breast-cancer>.
- [44] Larissa Vaz-Gonçalves, Louise Marquart-Wilson, Melinda M. Protani, Meghan T. Stephensen, Julie Moore, Michelle F. Morris, Jodi M. Saunus, and Marina M. Reeves. Capturing breast cancer subtypes in cancer registries: Insights into real-world incidence and survival. *Journal of Cancer Policy*, 44:100567, June 2025.
- [45] L. van Walle, J. Vandeven, C. Colpaert, FP. Duhoux, P. Neven, L. Van Eycken, and N. Van Damme. Incidence of breast cancer subtypes in Belgium: a population-based study. *Belgian Journal of Medical Oncology*, 14(6):263–73, 2020.
- [46] Yinghao Su, Ying Zheng, Wei Zheng, Kai Gu, Zhi Chen, Guoliang Li, Qiuyin Cai, Wei Lu, and Xiao Ou Shu. Distinct distribution and prognostic significance of molecular subtypes of breast cancer in Chinese women: a population-based cohort study. *BMC Cancer*, 11(1):292, July 2011.
- [47] V. S. P. K. Sankara Aditya Jayanthi, Asim Bikas Das, and Urmila Saxena. Grade-specific diagnostic and prognostic biomarkers in breast cancer. *Genomics*, 112(1):388–396, January 2020.
- [48] C. W. Elston and I. O. Ellis. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology*, 19(5):403–410, November 1991.
- [49] Stephen Fox, Valerie Speirs, and Abeer M. Shaaban. Male breast cancer: an update. *Virchows Archiv: An International Journal of Pathology*, 480(1):85–93, January 2022.
- [50] Jake Probert, John Broggio, Sarah C. Darby, David Dodwell, Paul McGale, Carolyn Taylor, and Kezia Gaitskell. Histological types of invasive breast cancer in 830,000 women diagnosed in England during 1988–2016. *The Journal of Pathology: Clinical Research*, 11(5):e70043, September 2025.
- [51] Prakash Pandit, Roshankumar Patil, Vijay Palwe, Sucheta Gandhe, Rahul Patil, and Rajnish Nagarkar. Prevalence of Molecular Subtypes of Breast Cancer: A Single Institutional Experience of 2062 Patients. *European Journal of Breast Health*, 16(1):39–43, January 2020.

- [52] SEER. Female Breast Cancer Subtypes - Cancer Stat Facts. NIH National Cancer Institute. Available from <https://seer.cancer.gov/statfacts/html/breast-subtypes.html>. Accessed 2/2/2026.
- [53] L. Ding, M. J. W. Greuter, I. Truyen, M. Goossens, H. De Schutter, G. H. de Bock, and G. Van Hal. Irregular screening participation increases advanced stage breast cancer at diagnosis: A population-based study. *Breast*, 65:61–66, October 2022.
- [54] Javier David Benitez Fuentes, Eileen Morgan, Alicia de Luna Aguilar, Allini Mafra, Richa Shah, Francesco Giusti, Jérôme Vignat, Ariana Znaor, Carina Musetti, Cheng-Har Yip, Liesbet Van Eycken, Elima Jedy-Agba, Marion Piñeros, and Isabelle Soerjomataram. Global Stage Distribution of Breast Cancer at Diagnosis: A Systematic Review and Meta-Analysis. *JAMA Oncology*, 10(1):71–78, January 2024.
- [55] Steven A. Narod. Breast cancer in young women. *Nature Reviews. Clinical Oncology*, 9(8):460–470, June 2012.
- [56] Ioannis Iakovou, Evanthia Giannoula, Areti Gkantaifi, Sofia Levva, and Savvas Frangos. Positron emission tomography in breast cancer: 18F-FDG and other radiopharmaceuticals. *European Journal of Hybrid Imaging*, 2, October 2018.
- [57] Bas Pouw, Linda J. de Wit-van der Veen, Marcel P. M. Stokkel, Claudette E. Loo, Marie-Jeanne T. F. D. Vrancken Peeters, and Renato A. Valdés Olmos. Heading toward radioactive seed localization in non-palpable breast cancer surgery? A meta-analysis. *Journal of Surgical Oncology*, 111(2):185–191, February 2015.
- [58] Francesco Giammarile, Sergi Vidal-Sicart, Diana Paez, Olivier Pellet, Estrada-Lobato Enrique, Miriam Mikhail-Lette, Olga Morozova, Navarro Marulanda Maria Camila, Rodríguez Sanchez Diana Ivonne, Roberto C. Delgado Bolton, Renato A. Valdés Olmos, and Giuliano Mariani. Sentinel Lymph Node Methods in Breast Cancer. *Seminars in Nuclear Medicine*, 52(5):551–560, September 2022.
- [59] C. C. Park, M. Mitsumori, A. Nixon, A. Recht, J. Connolly, R. Gelman, B. Silver, S. Hetelekidis, A. Abner, J. R. Harris, and S. J. Schnitt. Outcome at 8 years after breast-conserving surgery and radiation therapy for invasive breast cancer: influence of margin status and systemic therapy on local recurrence. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 18(8):1668–1675, April 2000.
- [60] Alberto Nunez, Veronica Jones, Katherine Schulz-Costello, and Daniel Schmolze. Accuracy of gross intraoperative margin assessment for breast cancer: experience since the SSO-ASTRO margin consensus guidelines. *Scientific Reports*, 10(1):17344, October 2020.

- [61] Nathan C Hall, Stephen P Povoski, Douglas A Murrey, Michael V Knopp, and Edward W Martin. Bringing advanced medical imaging into the operative arena could revolutionize the surgical care of cancer patients. *Expert Review of Medical Devices*, 5(6):663–667, November 2008.
- [62] M. A. Molina, S. Snell, D. Franceschi, M. Jorda, C. Gomez, F. L. Moffat, J. Powell, and E. Avisar. Breast specimen orientation. *Annals of Surgical Oncology*, 16(2):285–288, February 2009.
- [63] Ariella M. Altman, David D. Nguyen, Benjamin Johnson, Schelomo Marmor, Molly E. Klein, Todd M. Tuttle, and Jane Yuet Ching Hui. Intraoperative inking is superior to suture marking for specimen orientation in breast cancer. *The Breast Journal*, 26(4):661–667, April 2020.
- [64] Lee G. Wilke, Tomasz Czechura, Chih Wang, Brittany Lapin, Erik Liederbach, David P. Winchester, and Katharine Yao. Repeat Surgery After Breast Conservation for the Treatment of Stage 0 to II Breast Carcinoma: A Report From the National Cancer Data Base, 2004-2010. *JAMA Surgery*, 149(12):1296, December 2014.
- [65] Merisa L. Piper, Jasmine Wong, Kelly Fahrner-Scott, Cheryl Ewing, Michael Alvarado, Laura J. Esserman, and Rita A. Mukhtar. Success rates of re-excision after positive margins for invasive lobular carcinoma of the breast. *NPJ breast cancer*, 5:29, 2019.
- [66] R. A. Sakr, B. Poulet, G. J. Kaufman, C. Nos, and K. B. Clough. Clear margins for invasive lobular carcinoma: a surgical challenge. *European Journal of Surgical Oncology: The Journal of the European Society of Surgical Oncology and the British Association of Surgical Oncology*, 37(4):350–356, April 2011.
- [67] Lior Z. Braunstein, Jane E. Brock, Yu-Hui Chen, Linh Truong, Andrea L. Russo, Nils D. Arvold, and Jay R. Harris. Invasive lobular carcinoma of the breast: local recurrence after breast-conserving therapy by subtype approximation and surgical margin. *Breast Cancer Research and Treatment*, 149(2):555–564, January 2015.
- [68] Lindsey R. Fauveau, Tuoc N. Dao, Lucy B. Wallace, Mufaddal K. Mamawala, Ala Obaid, Anthony C. Waddimba, and Michael D. Grant. Positive surgical margins after breast-conserving surgery for ductal carcinoma in-situ: does histologic grade or estrogen receptor status matter? *Breast Cancer Research and Treatment*, 199(2):215–220, June 2023.
- [69] Sahej Dhak, Christopher Baliski, and Brendan Bakos. Factors influencing suboptimal pathologic margins and re-excision following breast conserving surgery for ductal carcinoma in-situ. *American Journal of Surgery*, 225(5):866–870, May 2023.

- [70] Gianluca Vanni, Marco Pellicciaro, Marco Materazzo, Valentina Marsella, Valeria Usai, Annalisa Noce, and Oreste Claudio Buonomo. Impact of cavity shave margins in patients with ductal carcinoma in situ undergoing conserving breast surgery. *Frontiers in Oncology*, 14:1403069, 2024.
- [71] Mila Trementosa Garcia, Bruna Salani Mota, Natalia Cardoso, Ana Luiza Cabrera Martimbiano, Marcos Desidério Ricci, Filomena Marino Carvalho, Rodrigo Gonçalves, José Maria Soares Junior, and José Roberto Filassi. Accuracy of frozen section in intraoperative margin assessment for breast-conserving surgery: A systematic review and meta-analysis. *PloS One*, 16(3):e0248768, 2021.
- [72] Gavin P. Dowling, Cian M. Hehir, Gordon R. Daly, Sandra Hembrecht, Stephen Keelan, Katie Giblin, Maen M. Alrawashdeh, Fiona Boland, and Arnold D.K. Hill. Diagnostic accuracy of intraoperative methods for margin assessment in breast cancer surgery: A systematic review & meta-analysis. *The Breast : Official Journal of the European Society of Mastology*, 76:103749, May 2024.
- [73] Irina Palimaru Manhoobi, Anne Bodilsen, Jasper Nijkamp, Anuj Pareek, Trine Tramm, Søren Redsted, and Peer Christiansen. Diagnostic accuracy of radiography, digital breast tomosynthesis, micro-CT and ultrasound for margin assessment during breast surgery: A systematic review and meta-analysis. *Academic Radiology*, 29(10):1560–1572, October 2022.
- [74] Dhurka Shanthakumar, Maria Leiloglou, Colm Kelliher, Ara Darzi, Daniel S. Elson, and Daniel R. Leff. A Comparison of Spectroscopy and Imaging Techniques Utilizing Spectrally Resolved Diffusely Reflected Light for Intraoperative Margin Assessment in Breast-Conserving Surgery: A Systematic Review and Meta-Analysis. *Cancers*, 15(11):2884, May 2023.
- [75] Mei-Huan Wang, Xiao Liu, Qian Wang, and Hua-Wei Zhang. Diagnosis accuracy of Raman spectroscopy in the diagnosis of breast cancer: a meta-analysis. *Analytical and Bioanalytical Chemistry*, 414(27):7911–7922, November 2022.
- [76] Yuqing Duan, Dingjie Guo, Xin Zhang, Linwei Lan, Hengyu Meng, Yashan Wang, Chuanying Sui, Zihan Qu, Guangliang He, Chunpeng Wang, and Xin Liu. Diagnostic accuracy of optical coherence tomography for margin assessment in breast-conserving surgery: A systematic review and meta-analysis. *Photodiagnosis and Photodynamic Therapy*, 43:103718, September 2023.
- [77] C. Elfgen, B. Papassotiropoulos, Z. Varga, L. Moskovszky, M. Nap, U. Güth, A. Baege, E. Amann, F. Chiesa, and C. Tausch. Comparative analysis of confocal microscopy on fresh breast core needle biopsies and conventional histology. *Diagnostic Pathology*, 14, June 2019.

- [78] J.M. Dixon, L. Renshaw, O. Young, D. Kulkarni, T. Saleem, M. Sarfaty, R. Sreenivasan, C. Kusnick, J. Thomas, and L.J. Williams. Intra-operative assessment of excised breast tumour margins using ClearEdge imaging device. *European Journal of Surgical Oncology (EJSO)*, 42(12):1834–1840, December 2016.
- [79] Electromagnetic (EM) Spectrum. Center for Science Education. Available from <https://scied.ucar.edu/learning-zone/earth-system/electromagnetic-spectrum>.
- [80] Fahim Ul-Hassan and Gary J Cook. PET/CT in oncology. *Clinical Medicine*, 12(4):368–372, 2012.
- [81] Koosha Paydary, Siavash Mehdizadeh Seraj, Mahdi Zirakchian Zadeh, Sahra Emamzadehfard, Sara Pourhassan Shamchi, Saeid Gholami, Thomas J. Werner, and Abass Alavi. The Evolving Role of FDG-PET/CT in the Diagnosis, Staging, and Treatment of Breast Cancer. *Molecular Imaging and Biology*, 21(1):1–10, 2019.
- [82] Eugene C. Lin and Abass Alavi. *PET and PET/CT: A Clinical Guide*. Thieme Medical Publishers, 3 edition, 2019.
- [83] Thorsten M. Buzug. *Computed Tomography: From Photon Statistics to Modern Cone-Beam CT*. Springer, 2008.
- [84] Michael E. Phelps. *PET: Physics, Instrumentation, and Scanners*. Springer, 2006.
- [85] Harjit Singh and Janet A. Neutze. *Radiology Fundamentals: Introduction to Imaging & Technology*. Springer, 2012.
- [86] Jorianne Boers, Erik F. J. de Vries, Andor W. J. M. Glaudemans, Geke A. P. Hospers, and Carolina P. Schröder. Application of PET Tracers in Molecular Imaging for Breast Cancer. *Current Oncology Reports*, 22(8):85, August 2020.
- [87] Jens M. Debacker, Vanessa Schelfhout, Lieve Brochez, David Creytens, Yves D’Asseler, Philippe Deron, Vincent Keereman, Koen Van de Vijver, Christian Vanhove, and Wouter Huvenne. High-Resolution 18F-FDG PET/CT for Assessing Three-Dimensional Intraoperative Margins Status in Malignancies of the Head and Neck, a Proof-of-Concept. *Journal of Clinical Medicine*, 10(16):3737, 2021.
- [88] Gary A. Ulaner. PET/CT for Patients With Breast Cancer: Where Is the Clinical Impact? *AJR. American journal of roentgenology*, 213(2):254–265, 2019.
- [89] Panli Li, Xiuying Wang, Chongrui Xu, Cheng Liu, Chaojie Zheng, Michael J Fulham, Dagan Feng, Lisheng Wang, Shaoli Song, and Gang Huang. 18F-FDG PET/CT radiomic predictors of pathologic complete response (pCR) to neoadjuvant chemotherapy in breast cancer

- patients. *European Journal of Nuclear Medicine and Molecular Imaging*, 47(5):1116–1126, May 2020.
- [90] Andrea Schaefer, Yoo Jin Kim, Stephanie Kremp, Sebastian Mai, Jochen Fleckenstein, Hendrik Bohnenberger, Hans-Joachim Schäfers, Jan-Martin Kuhnigk, Rainer M. Bohle, Christian Rübe, Carl-Martin Kirsch, and Aleksandar Grgic. PET-based delineation of tumour volumes in lung cancer: comparison with pathological findings. *European Journal of Nuclear Medicine and Molecular Imaging*, 40(8):1233–1244, August 2013.
- [91] Andrea Schaefer, Stephanie Kremp, Dirk Hellwig, Christian Rübe, Carl-Martin Kirsch, and Ursula Nestle. A contrast-oriented algorithm for FDG-PET-based delineation of tumour volumes for the radiotherapy of lung cancer: derivation from phantom measurements and validation in patient data. *European Journal of Nuclear Medicine and Molecular Imaging*, 35(11):1989–1999, November 2008.
- [92] Sarayut Janmahasatian, Stephen B. Duffull, Susan Ash, Leigh C. Ward, Nuala M. Byrne, and Bruce Green. Quantification of lean bodyweight. *Clinical Pharmacokinetics*, 44(10):1051–1065, 2005.
- [93] Freya Schnabel, Susan K. Boolbol, Mark Gittleman, Tami Karni, Lorraine Tafra, Sheldon Feldman, Alice Police, Neil B. Friedman, Scott Karlan, Dennis Holmes, Shawna C. Willey, Moshe Carmon, Kristen Fernandez, Stephanie Akbari, Jay Harness, Lisa Guerra, Thomas Frazier, Karen Lane, Rache M. Simmons, Alison Estabrook, and Tanir Allweis. A Randomized Prospective Study of Lumpectomy Margin Assessment with Use of MarginProbe in Patients with Nonpalpable Breast Malignancies. *Annals of Surgical Oncology*, 21(5):1589–1595, May 2014.
- [94] Adam M. Zysk, Kai Chen, Edward Gabrielson, Lorraine Tafra, Evelyn A. May Gonzalez, Joseph K. Canner, Eric B. Schneider, Andrew J. Cittadine, P. Scott Carney, Stephen A. Boppart, Kimiko Tsuchiya, Kristen Sawyer, and Lisa K. Jacobs. Intraoperative Assessment of Final Margins with a Handheld Optical Imaging Probe During Breast-Conserving Surgery May Reduce the Reoperation Rate: Results of a Multicenter Study. *Annals of Surgical Oncology*, 22(10):3356–3362, October 2015.
- [95] Trishan Panch, Peter Szolovits, and Rifat Atun. Artificial intelligence, machine learning and health systems. *Journal of Global Health*, 8(2):020303, 2018.
- [96] Milan Decuyper, Jens Maebe, Roel Van Holen, and Stefaan Vandenberghe. Artificial intelligence with deep learning in nuclear medicine and radiology. *EJNMMI Physics*, 8(1):81, December 2021.
- [97] Christian Janiesch, Patrick Zschech, and Kai Heinrich. Machine learning and deep learning. *Electronic Markets*, 31(3):685–695, September 2021.

- [98] Kunihiro Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, April 1980.
- [99] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier Non-linearities Improve Neural Network Acoustic Models. *International Conference on Machine Learning*, 2013.
- [100] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA, June 2016. IEEE.
- [101] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, January 1989.
- [102] Y. Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Proceedings of the USSR Academy of Sciences*, 1983.
- [103] Stuart Geman, Elie Bienenstock, and René Doursat. Neural Networks and the Bias/Variance Dilemma. *Neural Computation*, 4(1):1–58, January 1992.
- [104] Pierre-Henri Conze, Gustavo Andrade-Miranda, Vivek Kumar Singh, Vincent Jaouen, and Dimitris Visvikis. Current and Emerging Trends in Medical Image Segmentation With Deep Learning. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 7(6):545–569, July 2023.
- [105] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. *IEEE Computer Vision and Pattern Recognition*, 2009.
- [106] Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. *Nature Medicine*, 25(1):24–29, January 2019.
- [107] Andre Esteva, Katherine Chou, Serena Yeung, Nikhil Naik, Ali Madani, Ali Mottaghi, Yun Liu, Eric Topol, Jeff Dean, and Richard Socher. Deep learning-enabled medical computer vision. *NPJ Digital Medicine*, 4:5, January 2021.
- [108] Ihab S Mohamed. Detection and Tracking of Pallets using a Laser Rangefinder and Machine Learning Techniques. *Robotics*, page 77, 2017.
- [109] Siyuan Qiao, Huiyu Wang, Chenxi Liu, Wei Shen, and Alan Yuille. Micro-Batch Training with Batch-Channel Normalization and Weight Standardization, August 2020. arXiv:1903.10520 [cs].

-
- [110] Md. Eshmam Rayed, S. M. Sajibul Islam, Sadia Islam Niha, Jamin Rahman Jim, Md Mohsin Kabir, and M. F. Mridha. Deep learning for medical image segmentation: State-of-the-art advancements and challenges. *Informatics in Medicine Unlocked*, 47:101504, January 2024.
- [111] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Lecture Notes in Computer Science, pages 234–241, Cham, 2015. Springer International Publishing.
- [112] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation, June 2016. arXiv:1606.04797 [cs].
- [113] Eric Kerfoot, James Clough, Ilkay Oksuz, Jack Lee, Andrew P. King, and Julia A. Schnabel. Left-Ventricle Quantification Using Residual U-Net. In Mihaela Pop, Maxime Sermesant, Jichao Zhao, Shuo Li, Kristin McLeod, Alistair Young, Kawal Rhode, and Tommaso Mansi, editors, *Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges*, Lecture Notes in Computer Science, pages 371–380, Cham, 2019. Springer International Publishing.
- [114] Dominik Müller, Iñaki Soto-Rey, and Frank Kramer. Towards a guideline for evaluation metrics in medical image segmentation. *BMC Research Notes*, 15(1):210, June 2022.
- [115] Shruti Jadon. A survey of loss functions for semantic segmentation. In *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–7, October 2020.
- [116] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, January 2024.
- [117] Fabian Isensee, Paul F. Jäger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. Automated Design of Deep Learning Methods for Biomedical Image Segmentation. *Nature Methods*, 18(2):203–211, February 2021. arXiv:1904.08128 [cs].
- [118] Cédric Hémon, Blanche Texier, Caroline Lafond, Jean-Claude Nunes, and Anaïs Barateau. Towards trustworthy AI in radiotherapy: a comprehensive review of uncertainty-aware techniques. *Physics in Medicine and Biology*, 71(1), December 2025.
- [119] Valery Manokhin. *Practical Guide to Applied Conformal Prediction in Python*. Packt Publishing, 2023.

- [120] Grant Haskins, Uwe Kruger, and Pingkun Yan. Deep learning in medical image registration: a survey. *Machine Vision and Applications*, 31(1):8, January 2020.
- [121] Fatemehzahra Darzi and Thomas Bocklitz. A Review of Medical Image Registration for Different Modalities. *Bioengineering*, 11(8):786, August 2024.
- [122] Hans-Jonas Meyer, Andreas Wienke, and Alexey Surov. Associations between GLUT expression and SUV values derived from FDG-PET in different tumors—A systematic review and meta analysis. *PLOS ONE*, 14(6):eo217781, 2019.
- [123] F. Schiller, T. Fechter, C. Zamboglou, A. Chirindel, N. Salman, C.A. Jilg, V. Drendel, M. Werner, P.T. Meyer, A.-L. Grosu, and M. Mix. Comparison of PET/CT and whole-mount histopathology sections of the human prostate: a new strategy for voxel-wise evaluation. *EJNMMI Physics*, 4(1):21, 2017.
- [124] Tanuj Puri, Anastasia Chalkidou, Rhonda Henley-Smith, Arunabha Roy, Paul R. Barber, Teresa Guerrero-Urbano, Richard Oakley, Ricard Simo, Jean-Pierre Jeannon, Mark McGurk, Edward W. Odell, Michael J. O'Doherty, and Paul K. Marsden. A method for accurate spatial registration of PET images and histopathology slices. *EJNMMI Research*, 5(1):64, 2015.
- [125] Tanuj Puri, Anastasia Chalkidou, Arunabha Roy, Rhonda Henley-Smith, Paul R. Barber, Richard Oakley, Richard Simo, Jean-Pierre Jeannon, Mark McGurk, and Paul Marsden. A method for accurate spatial registration of PET images and histopathology slices. *Oncology News*, 11:10–12, March 2016.
- [126] Anando Sen, Patricia Troncoso, Aradhana Venkatesan, Mark D. Pagel, Jasper A. Nijkamp, Yulun He, Anne-Cecille Lesage, McKell Woodland, and Kristy K. Brock. Correlation of in-vivo imaging with histopathology: A review. *European Journal of Radiology*, 144:109964, 2021.
- [127] Lisanne L. de Boer, Esther Kho, Jasper Nijkamp, Koen K. Van de Vijver, Henricus J. C. M. Sterenborg, Leon C. ter Beek, and Theo J. M. Ruers. Method for coregistration of optical measurements of breast tissue with histopathology: the importance of accounting for tissue deformations. *Journal of Biomedical Optics*, 24(7):075002, 2019.
- [128] Jakob Unger, Tianchen Sun, Yi-Ling Chen, Jennifer E. Phipps, Richard J. Bold, Morgan A. Darrow, Kwan-Liu Ma, and Laura Marcu. Method for accurate registration of tissue autofluorescence imaging data with corresponding histology: a means for enhanced tumor margin assessment. *Journal of Biomedical Optics*, 23(1):015001, 2018.

- [129] Ashley M. Laughney, Venkataramanan Krishnaswamy, Elizabeth J. Rizzo, Mary C. Schwab, Richard J. Barth, Brian W. Pogue, Keith D. Paulsen, and Wendy A. Wells. Scatter spectroscopic imaging distinguishes between breast pathologies in tissues relevant to surgical margin assessment. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 18(22):6315–6325, 2012.
- [130] Stephen P Povoski, Nathan C Hall, Edward W Martin, and Michael J Walker. Multimodality approach of perioperative 18F-FDG PET/CT imaging, intraoperative 18F-FDG handheld gamma probe detection, and intraoperative ultrasound for tumor localization and verification of resection of all sites of hypermetabolic activity in a case of occult recurrent metastatic melanoma. *World Journal of Surgical Oncology*, 6(1):1, December 2008.
- [131] Nathan C Hall, Stephen P Povoski, Douglas A Murrey, Michael V Knopp, and Edward W Martin. Combined approach of perioperative 18F-FDG PET/CT imaging and intraoperative 18F-FDG handheld gamma probe detection for tumor localization and verification of complete tumor resection in breast cancer. *World Journal of Surgical Oncology*, 5(1):143, December 2007.
- [132] David E. Cohn, Nathan C. Hall, Stephen P. Povoski, Leigh G. Seamon, William B. Farrar, and Edward W. Martin. Novel perioperative imaging with 18F-FDG PET/CT and intraoperative 18F-FDG detection using a handheld gamma probe in recurrent ovarian cancer. *Gynecologic Oncology*, 110(2):152–157, August 2008.
- [133] Lorenzo Muraglia, Francesco Mattana, Laura Lavinia Travaini, Gennaro Musi, Emilio Bertani, Giuseppe Renne, Eleonora Pisa, Mahila Esmeralda Ferrari, Uberto Fumagalli Romario, Ottavio De Cobelli, Nicola Fusco, and Francesco Ceci. First Live-Experience Session with PET/CT Specimen Imager: A Pilot Analysis in Prostate Cancer and Neuroendocrine Tumor. *Biomedicines*, 11(2):645, February 2023.
- [134] Christopher Darr, Pedro Fragoso Costa, Theresa Kahl, Alexandros Moraitis, Jenna Engel, Mulham Al-Nader, Henning Reis, Jens Köllermann, Claudia Kesch, Ulrich Krafft, Tobias Maurer, Daniel Köhler, Susanne Klutmann, Fabian Falkenbach, Jens Kleesiek, Wolfgang P. Fendler, Boris A. Hadaschik, and Ken Herrmann. Intraoperative Molecular Positron Emission Tomography Imaging for Intraoperative Assessment of Radical Prostatectomy Specimens. *European Urology Open Science*, 54:28–32, 2023.
- [135] Jens M. Debacker, Luna Maris, Fleur Cordier, David Creytens, Philippe Deron, Benedicte Descamps, Yves D’Asseler, Kathia De Man, Vincent Keereman, Sasha Libbrecht, Vanessa Schelfhout, Koen

- Van De Vijver, Christian Vanhove, and Wouter Huvenne. Direct co-registration of [18F]FDG uptake and histopathology in surgically excised malignancies of the head and neck: a feasibility study. *European Journal of Nuclear Medicine and Molecular Imaging*, 50(7):2127–2139, June 2023.
- [136] Amit Agrawal, Nathan C. Hall, Matthew D. Ringel, Stephen P. Povoski, and Edward W. Martin. Combined Use of Perioperative TSH-Stimulated 18F-FDG PET/CT Imaging and Gamma Probe Radioguided Surgery to Localize and Verify Resection of Iodine Scan-Negative Recurrent Thyroid Carcinoma. *The Laryngoscope*, 118(12):2190–2194, December 2008.
- [137] Luna Maris, Kathia De Man, Filip Gryspeerdt, Ann Hoorens, Vincent Keereman, Anneleen Kiekens, Bliede Van den Broeck, Leen Van de Sande, Christian Vanhove, and Frederik Berrevoet. Abstract: 18F-FDG-PET-CT specimen imaging for perioperative visualization of pancreatic adenocarcinoma: a proof-of-concept study. ESSO41 - Online Congress Portal - Presentation number: P292, 2022.
- [138] Marc J. Gollub, Timothy J. Akhurst, Matthew J. Williamson, Jinru Shia, John L. Humm, W. Douglas Wong, Philip B. Paty, Jose G. Guillem, Martin R. Weiser, Larissa K. F. Temple, Lawrence T. Dauer, Suresh C. Jhanwar, Rachel E. Kronman, Carolina V. Montalvo, Allison R. Miller, Steven M. Larson, and Alexander R. Margulis. Feasibility of ex Vivo FDG PET of the Colon. *Radiology*, 252(1):232–239, July 2009.
- [139] Angela Van Baardwijk, Geert Bosmans, Robert Jan Van Suylen, Marinus Van Kroonenburgh, Monique Hochstenbag, Gijs Geskes, Philippe Lambin, and Dirk De Ruyscher. Correlation of intra-tumour heterogeneity on 18F-FDG PET with pathologic features in non-small cell lung cancer: A feasibility study. *Radiotherapy and Oncology*, 87(1):55–58, April 2008.
- [140] Susan D. Moffatt-Bruce, Stephen P. Povoski, Suhail Sharif, Nathan C. Hall, Patrick Ross, Morgan A. Johnson, and Edward W. Martin. A Novel Approach to Positron Emission Tomography in Lung Cancer. *The Annals of Thoracic Surgery*, 86(4):1355–1357, October 2008.
- [141] Bawinile Hadebe, Lerwine Harry, Tasmeera Ebrahim, Venesen Pillay, and Mariza Vorster. The Role of PET/CT in Breast Cancer. *Diagnostics*, 13(4):597, 2023. Number: 4.
- [142] Sang Kyu Yang, Nariya Cho, and Woo Kyung Moon. The Role of PET/CT for Evaluating Breast Cancer. *Korean Journal of Radiology*, 8(5):429–437, 2007.
- [143] Takaaki Fujii, Reina Yajima, Sasagu Kurozumi, Toru Higuchi, Sayaka Obayashi, Hideaki Tokiniwa, Rin Nagaoka, Daisuke Takata, Jun Horiguchi, and Hiroyuki Kuwano. Clinical Significance of 18F-FDG-PET

- in Invasive Lobular Carcinoma. *Anticancer Research*, 36(10):5481–5485, 2016.
- [144] Jonas Albers, Angelika Svetlove, Justus Alves, Alexander Kraupner, Francesca di Lillo, M. Andrea Markus, Giuliana Tromba, Frauke Alves, and Christian Dullin. Elastic transformation of histological slices allows precise co-registration with microCT data sets for a refined virtual histology approach. *Scientific Reports*, 11(1):10846, 2021. Number: 1.
- [145] Kasper Marstal, Floris Berendsen, Marius Staring, and Stefan Klein. SimpleElastix: A User-Friendly, Multi-lingual Library for Medical Image Registration. *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 574–582, 2016. Conference Name: 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) ISBN: 9781509014378 Place: Las Vegas, NV, USA.
- [146] Itseez. Open Source Computer Vision Library, 2015.
- [147] Andreas Loening and Shweta Gambhir. AMIDE: A Free Software Tool for Multimodality Medical Image Analysis. *Molecular Imaging*, 2, 2003.
- [148] Peter Bankhead, Maurice B. Loughrey, José A. Fernández, Yvonne Dombrowski, Darragh G. McArt, Philip D. Dunne, Stephen McQuaid, Ronan T. Gray, Liam J. Murray, Helen G. Coleman, Jacqueline A. James, Manuel Salto-Tellez, and Peter W. Hamilton. QuPath: Open source software for digital pathology image analysis. *Scientific Reports*, 7(1):16878, 2017. Number: 1.
- [149] Aisheng Dong, Yang Wang, Jianping Lu, and Changjing Zuo. Spectrum of the Breast Lesions With Increased 18F-FDG Uptake on PET/CT. *Clinical Nuclear Medicine*, 41(7):543–557, 2016.
- [150] Asako Azuma, Mitsuhiro Tozaki, Kensuke Ito, Eisuke Fukuma, Tomoko Tanaka, and Toshihiro O’uchi. Ductal carcinoma in situ: correlation between FDG-PET/CT and histopathology. *Radiation Medicine*, 26(8):488–493, 2008.
- [151] Ashraf Anas Zytoon, Koji Murakami, Mohamed Ramdan El-Kholy, Emad El-Shorbagy, and Osama Ebied. Breast cancer with low FDG uptake: Characterization by means of dual-time point FDG-PET/CT. *European Journal of Radiology*, 70(3):530–538, 2009.
- [152] Paul E. Kinahan and James W. Fletcher. PET/CT Standardized Uptake Values (SUVs) in Clinical Practice and Assessing Response to Therapy. *Seminars in ultrasound, CT, and MR*, 31(6):496–505, December 2010.

- [153] T. Sengoz, Y. A. Karakaya, A. Gültekin, O. Yaylali, H. Senol, and D. Yuksel. Relationships of 18F-FDG uptake by primary tumors with prognostic factors and molecular subtype in ductal breast cancer. *Revista Espanola De Medicina Nuclear E Imagen Molecular*, pages S2253–654X(20)30211–0, 2021.
- [154] Tomoko Hirakata, Takaaki Fujii, Sasagu Kurozumi, Ayaka Katayama, Chikako Honda, Keiko Yanai, Shoko Tokuda, Yuko Nakazawa, Sayaka Obayashi, Reina Yajima, Kyoichi Kaira, Tetsunari Oyama, and Ken Shirabe. FDG uptake reflects breast cancer immunological features: the PD-L1 expression and degree of TILs in primary breast cancer. *Breast Cancer Research and Treatment*, 181(2):331–338, 2020.
- [155] Soeun Park, Eun-ki Min, Soong June Bae, Chihwan Cha, Dooreh Kim, Janghee Lee, Yoon Jin Cha, Sung Gwe Ahn, and Joon Jeong. Relationship of the standard uptake value of 18F-FDG-PET-CT with tumor-infiltrating lymphocytes in breast tumors measuring ≥ 1 cm. *Scientific Reports*, 11(1):12046, June 2021. Number: 1.
- [156] Mei Tian, Xuexin He, Chentao Jin, Xiao He, Shuang Wu, Rui Zhou, Xiaohui Zhang, Kai Zhang, Weizhong Gu, Jing Wang, and Hong Zhang. Transpathology: molecular imaging-based pathology. *European Journal of Nuclear Medicine and Molecular Imaging*, 48(8):2338–2350, 2021.
- [157] Syed Ahmed Taqi, Syed Abdus Sami, Lateef Begum Sami, and Syed Ahmed Zaki. A review of artifacts in histopathology. *Journal of Oral and Maxillofacial Pathology : JOMFP*, 22(2):279, 2018.
- [158] Mirabela Rusu, Prabhakar Rajiah, Robert Gilkeson, Michael Yang, Christopher Donatelli, Rajat Thawani, Frank J. Jacono, Philip Linden, and Anant Madabhushi. 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. *European Radiology*, 27(10):4209–4217, October 2017.
- [159] Mirabela Rusu, Wei Shao, Christian A. Kunder, Jeffrey B. Wang, Simon J. C. Soerensen, Nikola C. Teslovich, Rewa R. Sood, Leo C. Chen, Richard E. Fan, Pejman Ghanouni, James D. Brooks, and Geoffrey A. Sonn. Registration of presurgical MRI and histopathology images from radical prostatectomy via RAPSODI. *Medical Physics*, 47(9):4177–4188, 2020. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/mp.14337>.
- [160] Nobuyuki Otsu. A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, January 1979. Conference Name: IEEE Transactions on Systems, Man, and Cybernetics.

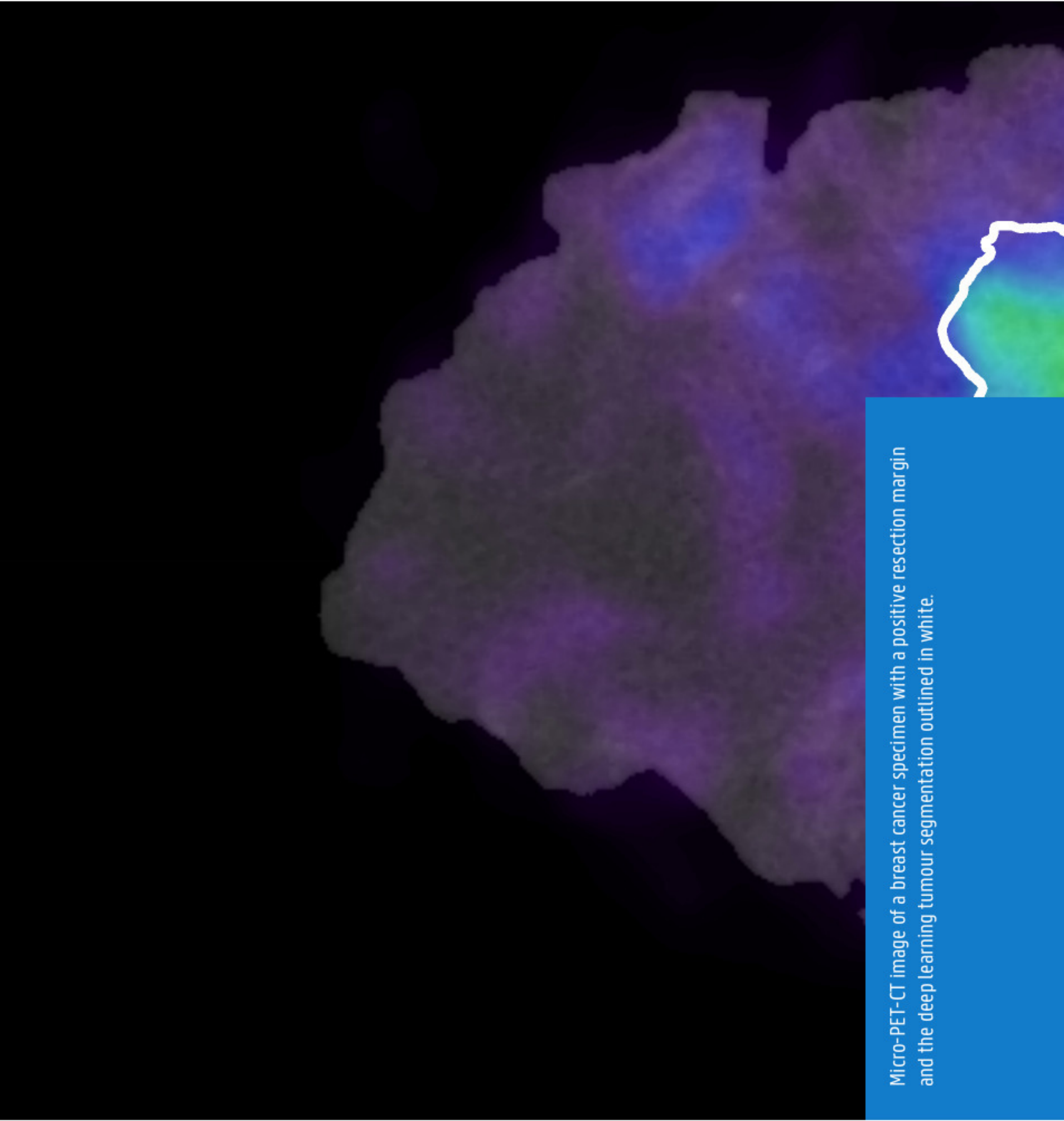
- [161] Nahian Siddique, Sidike Paheding, Colin P. Elkin, and Vijay Devabhaktuni. U-Net and Its Variants for Medical Image Segmentation: A Review of Theory and Applications. *IEEE Access*, 9:82031–82057, 2021. Conference Name: IEEE Access.
- [162] M. Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, Vishwesh Nath, Yufan He, Ziyue Xu, Ali Hatamizadeh, Andriy Myronenko, Wentao Zhu, Yun Liu, Mingxin Zheng, Yucheng Tang, Isaac Yang, Michael Zephyr, Behrooz Hashemian, Sachidanand Alle, Mohammad Zalbagi Darestani, Charlie Budd, Marc Modat, Tom Vercauteren, Guotai Wang, Yiwen Li, Yipeng Hu, Yunguan Fu, Benjamin Gorman, Hans Johnson, Brad Genereaux, Barbaros S. Erdal, Vikash Gupta, Andres Diaz-Pinto, Andre Dourson, Lena Maier-Hein, Paul F. Jaeger, Michael Baumgartner, Jayashree Kalpathy-Cramer, Mona Flores, Justin Kirby, Lee A. D. Cooper, Holger R. Roth, Daguang Xu, David Bericat, Ralf Floca, S. Kevin Zhou, Haris Shuaib, Keyvan Farahani, Klaus H. Maier-Hein, Stephen Aylward, Prerna Dogra, Sebastien Ourselin, and Andrew Feng. MONAI: An open-source framework for deep learning in healthcare, November 2022. arXiv:2211.02701 [cs].
- [163] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, volume 32, pages 8026–8037. Curran Associates, Inc., 2019.
- [164] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19*, pages 2623–2631, New York, NY, USA, July 2019. Association for Computing Machinery.
- [165] Susan Mercieca, José Belderbos, Judith van Loon, Kenneth Gilhuijs, Peter Julyan, and Marcel van Herk. Comparison of SUVmax and SUVpeak based segmentation to determine primary lung tumour volume on FDG PET-CT correlated with pathology data. *Radiotherapy and Oncology*, 129(2):227–233, November 2018.
- [166] Moon Kim, Robert Seifert, Jana Fragemann, David Kersting, Jacob Murray, Frederic Jonske, Kelsey L. Pomykala, Jan Egger, Wolfgang P. Fendler, Ken Herrmann, and Jens Kleesiek. 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. *European Journal of Nuclear Medicine and Molecular Imaging*, 50(7):2196–2209, 2023.
- [167] Yuvnik Trada, Peter Lin, Mark T. Lee, Michael G. Jameson, Phillip Chlap, Paul Keall, Daniel Moses, and Allan Fowler. Impact of tumour region of interest delineation method for mid-treatment FDG-PET response prediction in head and neck squamous cell carcinoma undergoing radiotherapy. *Quantitative Imaging in Medicine and Surgery*, 13(5):2822–2836, May 2023.
- [168] Habib Zaidi and Issam El Naqa. PET-guided delineation of radiation therapy treatment volumes: a survey of image segmentation techniques. *European Journal of Nuclear Medicine and Molecular Imaging*, 37(11):2165–2187, November 2010.
- [169] Mathieu Hatt, John A. Lee, Charles R. Schmidtlein, Issam El Naqa, Curtis Caldwell, Elisabetta De Bernardi, Wei Lu, Shiva Das, Xavier Geets, Vincent Gregoire, Robert Jeraj, Michael P. MacManus, Osama R. Mawlawi, Ursula Nestle, Andrei B. Pugachev, Heiko Schöder, Tony Shepherd, Emiliano Spezi, Dimitris Visvikis, Habib Zaidi, and Assen S. Kirov. Classification and evaluation strategies of auto-segmentation approaches for PET: Report of AAPM task group No. 211. *Medical Physics*, 44(6):e1–e42, June 2017.
- [170] Dinusha Veluponnar, Lisanne L. de Boer, Freija Geldof, Lynn-Jade S. Jong, Marcos Da Silva Guimaraes, Marie-Jeanne T. F. D. Vrancken Peeters, Frederieke van Duijnhoven, Theo Ruers, and Behdad Dashtbozorg. Toward Intraoperative Margin Assessment Using a Deep Learning-Based Approach for Automatic Tumor Segmentation in Breast Lumpectomy Ultrasound Images. *Cancers*, 15(6):1652, March 2023.
- [171] Chris Yeung, Tamas Ungi, Zoe Hu, Amoon Jamzad, Martin Kaufmann, Ross Walker, Shaila Merchant, Cecil Jay Engel, Doris Jabs, John Rudan, Parvin Mousavi, and Gabor Fichtinger. From quantitative metrics to clinical success: assessing the utility of deep learning for tumor segmentation in breast surgery. *International Journal of Computer Assisted Radiology and Surgery*, 19(6):1193–1201, June 2024.
- [172] Wei-Chung Shia, Yu-Hsun Kuo, Fang-Rong Hsu, Joseph Lin, Wen-Pei Wu, Hwa-Koon Wu, Wei-Cheng Yeh, and Dar-Ren Chen. Evaluating the Margins of Breast Cancer Tumors by Using Digital Breast Tomosynthesis with Deep Learning: A Preliminary Assessment. *Diagnostics*, 14(10):1032, January 2024. Number: 10.
- [173] Esther Kho, Behdad Dashtbozorg, Lisanne L. de Boer, Koen K. Van de Vijver, Henricus J. C. M. Sterenborg, and Theo J. M. Ruers. Broadband hyperspectral imaging for breast tumor detection using spectral

- and spatial information. *Biomedical Optics Express*, 10(9):4496, September 2019.
- [174] Mohamed Amgad, Habiba Elfandy, Hagar Hussein, Lamees A Atteya, Mai A T Elsebaie, Lamia S Abo Elnasr, Rokia A Sakr, Hazem S E Salem, Ahmed F Ismail, Anas M Saad, Joumana Ahmed, Maha A T Elsebaie, Mustafijur Rahman, Inas A Ruhban, Nada M Elgazar, Yahya Alagha, Mohamed H Osman, Ahmed M Alhusseiny, Mariam M Khalaf, Abo-Alela F Younes, Ali Abdulkarim, Duaa M Younes, Ahmed M Gadallah, Ahmad M Elkashash, Salma Y Fala, Basma M Zaki, Jonathan Beezley, Deepak R Chittajallu, David Manthey, David A Gutman, and Lee A D Cooper. Structured crowdsourcing enables convolutional segmentation of histology images. *Bioinformatics*, 35(18):3461–3467, September 2019.
- [175] Luna Maris, Menekse Göker, Jens M. Debacker, Kathia De Man, Bliede Van den Broeck, Jo Van Dorpe, Koen Van de Vijver, Vincent Keereman, and Christian Vanhove. Method for co-registration of high-resolution specimen PET-CT with histopathology to improve insight into radiotracer distributions. *EJNMMI Physics*, 11(1):85, October 2024.
- [176] Susan C. Lester, Shikha Bose, Yunn-Yi Chen, James L. Connolly, Monica E. de Baca, Patrick L. Fitzgibbons, Daniel F. Hayes, Celina Kleer, Frances P. O'Malley, David L. Page, Barbara L. Smith, Lee K. Tan, Donald L. Weaver, Eric Winer, and Members of the Cancer Committee, College of American Pathologists. Protocol for the examination of specimens from patients with invasive carcinoma of the breast. *Archives of Pathology & Laboratory Medicine*, 133(10):1515–1538, October 2009.
- [177] Susan C. Lester, Shikha Bose, Yunn-Yi Chen, James L. Connolly, Monica E. de Baca, Patrick L. Fitzgibbons, Daniel F. Hayes, Celina Kleer, Frances P. O'Malley, David L. Page, Barbara L. Smith, Donald L. Weaver, Eric Winer, and Members of the Cancer Committee, College of American Pathologists. Protocol for the examination of specimens from patients with ductal carcinoma in situ of the breast. *Archives of Pathology & Laboratory Medicine*, 133(1):15–25, January 2009.
- [178] Perrin E. Romine, Lanell M. Peterson, Brenda F. Kurland, Darrin W. Byrd, Alena Novakova-Jiresova, Mark Muzi, Jennifer M. Specht, Robert K. Doot, Jeanne M. Link, Kenneth A. Krohn, Paul E. Kinahan, David A. Mankoff, and Hannah M. Linden. 18F-fluorodeoxyglucose (FDG) PET or 18F-fluorothymidine (FLT) PET to assess early response to aromatase inhibitors (AI) in women with ER+ operable breast cancer in a window-of-opportunity study. *Breast Cancer Research : BCR*, 23:88, 2021.
- [179] Chikako Sekine, Nachiko Uchiyama, Chikashi Watase, Takeshi Murata, Sho Shiino, Kenjiro Jimbo, Eriko Iwamoto, Shin Takayama, Hiroaki

- Kurihara, Kaishi Satomi, Masayuki Yoshida, Takayuki Kinoshita, and Akihiko Suto. Preliminary experiences of PET/MRI in predicting complete response in patients with breast cancer treated with neoadjuvant chemotherapy. *Molecular and Clinical Oncology*, 16(2):50, February 2022.
- [180] YA Masannat, A Agrawal, L Maraqa, M Fuller, SK Down, SSK Tang, D Pang, M Kontos, L Romics, and SD Heys. Multifocal and multicentric breast cancer, is it time to think again? *Annals of The Royal College of Surgeons of England*, 102(1):62–66, January 2020.
- [181] Alireza Mehrtash, William M. Wells, Clare M. Tempany, Purang Abolmaesumi, and Tina Kapur. Confidence Calibration and Predictive Uncertainty Estimation for Deep Medical Image Segmentation. *IEEE transactions on medical imaging*, 39(12):3868–3878, December 2020.
- [182] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer International Publishing, Cham, 2022.
- [183] Vladimir Vovk and Ivan Petej. Venn-Abers predictors. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, pages 829–838, 2014.
- [184] Vladimir Vovk, Ivan Petej, and Valentina Fedorova. Large-scale probabilistic predictors with and without guarantees of validity. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.
- [185] Ivan Petej. venn-abers: Venn-Abers calibration for probabilistic classifiers, March 2026. original-date: 2023-07-16T20:32:12Z.
- [186] Samuel Davenport. Conformal confidence sets for biomedical image segmentation. *ICLR 2025*, 2025. Version Number: 2.
- [187] Luca Mossina and Corentin Friedrich. Conformal Prediction for Image Segmentation Using Morphological Prediction Sets. In James C. Gee, Daniel C. Alexander, Jaesung Hong, Juan Eugenio Iglesias, Carole H. Sudre, Archana Venkataraman, Polina Golland, Jong Hyo Kim, and Jinah Park, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, pages 78–88, Cham, 2026. Springer Nature Switzerland.
- [188] Anastasios N. Angelopoulos, Rina Foygel Barber, and Stephen Bates. Theoretical Foundations of Conformal Prediction, March 2026. arXiv:2411.11824 [math].
- [189] Alessia De Biase, Liv Ziegfeld, Nanna Maria Sijtsema, Roel Steenbakkers, Robin Wijsman, Lisanne V. van Dijk, Johannes A. Langendijk, Fokie Cnossen, and Peter van Ooijen. Probability maps for deep

learning-based head and neck tumor segmentation: Graphical User Interface design and test. *Computers in Biology and Medicine*, 177:108675, July 2024.



Micro-PET-CT image of a breast cancer specimen with a positive resection margin and the deep learning tumour segmentation outlined in white.